

Bandit Algorithms

Tor Lattimore and Csaba Szepesvári

Draft of Friday 27th July, 2018

Revision: 1014M

Contents

<i>Preface</i>	<i>page</i> 1
<i>Notation</i>	3
Part I Bandits, Probability and Concentration	7
1 Introduction	8
1.1 The language of bandits	9
1.2 Applications	13
1.3 Bibliographic remarks	16
2 Foundations of Probability (†)	18
2.1 Probability spaces and random elements	18
2.2 σ -algebras and knowledge	24
2.3 Conditional probabilities	26
2.4 Independence	28
2.5 Integration and expectation	29
2.6 Conditional expectation	32
2.7 Notes	35
2.8 Bibliographic remarks	39
2.9 Exercises	40
3 Stochastic Processes and Markov Chains (†)	44
3.1 Stochastic processes	45
3.2 Markov chains	45
3.3 Martingales and stopping times	47
3.4 Notes	48
3.5 Bibliographic remarks	49
3.6 Exercises	49
4 Finite-Armed Stochastic Bandits	51
4.1 The learning objective	52
4.2 The regret	54
4.3 Decomposing the regret	56
4.4 The canonical bandit model (†)	57

4.5	Notes	59
4.6	Bibliographical remarks	61
4.7	Exercises	62
5	Concentration of Measure	66
5.1	The inequalities of Markov and Chebyshev	67
5.2	The Cramer-Chernoff method and subgaussian random variables	68
5.3	Notes	71
5.4	Bibliographical remarks	74
5.5	Exercises	75
Part II	Stochastic Bandits with Finitely Many Arms	82
6	The Explore-then-Commit Algorithm	84
6.1	Notes	88
6.2	Bibliographical remarks	88
6.3	Exercises	89
7	The Upper Confidence Bound Algorithm	95
7.1	Notes	103
7.2	Bibliographical remarks	103
7.3	Exercises	104
8	The Upper Confidence Bound Algorithm: Asymptotic Optimality	110
8.1	Notes	113
8.2	Bibliographic remarks	113
8.3	Exercises	114
9	The Upper Confidence Bound Algorithm: Minimax Optimality (\dagger)	116
9.1	Notes	119
9.2	Bibliographic remarks	121
9.3	Exercises	121
10	The Upper Confidence Bound Algorithm: Bernoulli Noise (\dagger)	125
10.1	Concentration for sums of Bernoulli random variables	125
10.2	The KL-UCB algorithm	128
10.3	Notes	132
10.4	Bibliographic remarks	133
10.5	Exercises	134
Part III	Adversarial Bandits with Finitely Many Arms	137
11	The Exp3 Algorithm	141
11.1	Importance-weighted estimators	143

11.2	The Exp3 algorithm	145
11.3	Regret analysis	146
11.4	Notes	150
11.5	Bibliographic remarks	152
11.6	Exercises	152
12	The Exp3-IX Algorithm	157
12.1	Regret analysis	158
12.2	Notes	162
12.3	Bibliographic remarks	163
12.4	Exercises	164
Part IV	Lower Bounds for Bandits with Finitely Many Arms	167
13	Lower Bounds: Basic Ideas	170
13.1	Notes	173
13.2	Exercises	174
14	Foundations of Information Theory (†)	175
14.1	The relative entropy	177
14.2	Notes	181
14.3	Bibliographic remarks	183
14.4	Exercises	183
15	Minimax Lower Bounds	185
15.1	Relative entropy between bandits	185
15.2	Minimax lower bounds	186
15.3	Notes	188
15.4	Bibliographic remarks	189
15.5	Exercises	189
16	Instance Dependent Lower Bounds	192
16.1	Asymptotic bounds	193
16.2	Finite-time bounds	195
16.3	Notes	196
16.4	Bibliographic remarks	197
16.5	Exercises	197
17	High Probability Lower Bounds	200
17.1	Stochastic bandits	201
17.2	Adversarial bandits	203
17.3	Notes	205
17.4	Bibliographic remarks	205
17.5	Exercises	205

Part V	Contextual and Linear Bandits	206
18	Contextual Bandits	208
18.1	Contextual bandits: one bandit per context	208
18.2	Bandits with expert advice	210
18.3	Can it go higher? Exp4	213
18.4	Regret analysis	213
18.5	Notes	216
18.6	Bibliographic remarks	218
18.7	Exercises	219
19	Stochastic Linear Bandits	221
19.1	Stochastic contextual bandits	221
19.2	Stochastic linear bandits	223
19.3	Regret analysis	225
19.4	Notes	227
19.5	Bibliographic remarks	229
19.6	Exercises	229
20	Confidence Bounds for Least Squares Estimators	231
20.1	Martingale noise and Laplace's method	233
20.2	Notes	238
20.3	Bibliographic remarks	238
20.4	Exercises	238
21	Optimal Design for Least Squares Estimators	241
21.1	Proof of Kiefer–Wolfowitz (\dagger)	242
21.2	Minimum volume ellipsoids and John's theorem (\dagger)	243
21.3	Notes	245
21.4	Bibliographic remarks	246
21.5	Exercises	246
22	Stochastic Linear Bandits with Finitely Many Arms	247
22.1	Bibliographic remarks	248
22.2	Exercises	249
23	Stochastic Linear Bandits with Sparsity	250
23.1	Sparse linear stochastic bandits	250
23.2	Elimination on the hypercube	251
23.3	Proof of technical lemma	255
23.4	UCB with sparsity	256
23.5	Online to confidence set conversion	256
23.6	Sparse online linear prediction	259
23.7	Notes	260

23.8	Bibliographical Remarks	260
23.9	Exercises	261
24	Minimax Lower Bounds for Stochastic Linear Bandits	263
24.1	Hypercube	264
24.2	Sphere	265
24.3	Sparse parameter vectors	266
24.4	Unrealizable case	267
24.5	Notes	269
24.6	Bibliographic remarks	269
24.7	Exercises	270
25	Asymptotic Lower Bounds for Stochastic Linear Bandits	271
25.1	Clouds looming for optimism	274
25.2	Notes	276
25.3	Bibliographic remarks	277
25.4	Exercises	277
Part VI	Adversarial Linear Bandits	278
26	Foundations of Convex Analysis (†)	280
26.1	Convex sets and functions	280
26.2	Jensen's inequality	281
26.3	Bregman divergence	282
26.4	Legendre functions	283
26.5	Optimization	284
26.6	Projections	285
26.7	Notes	286
26.8	Bibliographic remarks	287
26.9	Exercises	287
27	Exp3 for Adversarial Linear Bandits	289
27.1	Exponential weights for linear bandits	289
27.2	Regret analysis	291
27.3	Continuous exponential weights	292
27.4	Notes	293
27.5	Bibliographic remarks	294
27.6	Exercises	294
28	Follow the Regularized Leader and Mirror Descent	296
28.1	Online linear optimization	296
28.2	Regret analysis	299
28.3	Online learning for bandits	303
28.4	The unit ball	304

28.5	Notes	306
28.6	Bibliographic remarks	308
28.7	Exercises	309
29	The Relation Between Adversarial and Stochastic Linear Bandits	313
29.1	Reducing stochastic linear bandits to adversarial linear bandits	314
29.2	Stochastic linear bandits with parameter noise	316
29.3	Notes	317
29.4	Bibliographic remarks	318
29.5	Exercises	318
Part VII	Other Topics	320
30	Combinatorial Bandits	324
30.1	Applications	324
30.2	Bandits	326
30.3	Semibandits	326
30.4	Follow the perturbed leader	328
30.5	Notes	333
30.6	Bibliographic remarks	334
30.7	Exercises	335
31	Non-Stationary Bandits	337
31.1	Adversarial bandits	337
31.2	Stochastic bandits	340
31.3	Notes	342
31.4	Bibliographic remarks	343
31.5	Exercises	344
32	Ranking	345
32.1	Click models	346
32.2	Policy	348
32.3	Regret analysis	350
32.4	Notes	354
32.5	Bibliographic remarks	356
32.6	Exercises	357
33	Pure Exploration	359
33.1	Simple regret	359
33.2	Best arm identification	361
33.3	Best arm identification with a budget	366
33.4	Notes	367
33.5	Bibliographical remarks	368
33.6	Exercises	369

34	Bayesian Methods	372
34.1	Bayesian optimal regret for finite-armed bandits	373
34.2	Bayesian learning (\dagger)	374
34.3	Conjugate priors and the exponential family (\dagger)	377
34.4	Bayesian learning and bandits	379
34.5	One-armed bandits	381
34.6	Gittins index	386
34.7	Computing the Gittins index	389
34.8	Notes	391
34.9	Bibliographical remarks	392
34.10	Exercises	393
35	Thompson Sampling	394
35.1	Finite-armed bandits	395
35.2	Linear bandits	400
35.3	Information theoretic analysis	402
35.4	Notes	405
35.5	Bibliographic remarks	407
35.6	Exercises	408
Part VIII Beyond Bandits		410
36	Partial Monitoring	411
36.1	Finite adversarial partial monitoring problems	412
36.2	The structure of partial monitoring	414
36.3	Classification of finite adversarial partial monitoring	418
36.4	Lower bounds	418
36.5	Policy for easy games	422
36.6	Upper bound for easy games	426
36.7	Proof of the classification theorem	430
36.8	Notes	431
36.9	Bibliographical remarks	434
36.10	Exercises	435
37	Markov Decision Processes	438
37.1	Problem setup	438
37.2	Optimal policies and the Bellman optimality equation	442
37.3	Finding an optimal policy (\dagger)	445
37.4	Learning in Markov decision processes	447
37.5	Upper confidence bounds for reinforcement learning	448
37.6	Proof of upper bound	451
37.7	Proof of lower bound	454
37.8	Notes	457
37.9	Bibliographical remarks	460

37.10 Exercises	462
Appendix A Bibliography	471
Index	500
<i>Solutions to Selected Exercises</i>	1

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 specialized literature by themselves, and to adapt existing algorithms to their applications. This latter point is important. As Tolstoy might have written, “problems in theory are all alike; every application is different”. A practitioner seeking to apply a bandit algorithm must 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 (mostly one-dimensional) and some linear algebra. The book uses the notation of measure-theoretic probability theory, but does not rely on any deep results. In addition, we introduce the notation and carefully (and hopefully intuitively) explain it along with the basic results that we need. This chapter is unusual for an introduction to measure theory in that it emphasizes the reasons to use σ -algebras beyond the standard technical justifications. We hope this will go some way to convince the reader that measure theory is an important and intuitive tool and not merely a weapon for rejecting papers written by non-mathematicians. 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 chapters contain content that is not really about bandits, such as convex analysis, information theory and probability. These chapters can

be skipped by knowledgeable readers, or otherwise referred to when necessary. They are marked with a (\dagger). Most chapters end with a list of notes and exercises. These are intended to deepen intuition and highlight the connections between various subsections and the literature. Following this preface there is a table of notation.

Thanks

Of course we will have many people to thank.

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.

 Something important or a warning to the reader.

 Hints and tips.

 For many algorithms we present simple experimental results showing their behaviour. These sections are marked with a beaker.

Basics

The sets of real and natural numbers are denoted by \mathbb{R} and \mathbb{N} , respectively, with $\mathbb{N} = \{0, 1, 2, 3, \dots\}$. We define $[n] = \{1, 2, 3, \dots, n\}$. The floor and ceiling functions of $x \in \mathbb{R}$ are $\lfloor x \rfloor$ and $\lceil x \rceil$, while $(x)^+ = \max(x, 0)$. For $p \geq 1$ the p -norm of vector $x \in \mathbb{R}^d$ is $\|x\|_p = (\sum_{i=1}^d x_i^p)^{1/p}$. The d -dimensional simplex embedded in \mathbb{R}^{d+1} is $\mathcal{P}_d = \{x \in [0, 1]^{d+1} : \|x\|_1 = 1\}$. A function $v : [n] \rightarrow \mathbb{R}$ is often associated with a vector using the same symbol $v \in \mathbb{R}^n$ which is defined in the natural way: $v_i = v(i)$.

Distributions and probability

The notation for probability and measure theory is introduced in Chapter 2. Those familiar with measure-theoretic probability can safely skip this chapter as the notation is consistent with the standard [Kallenberg, 2002, Billingsley, 2008]. We mention here that $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2 and that $\mathcal{B}(p)$ is the Bernoulli distribution with bias p . The set of all distributions on finite or countable set A is $\mathcal{P}(A)$. We use $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ for the expectation and variance operators respectively. By and large

the underlying measure is omitted from the notation, but sometimes we find it necessary to write $\mathbb{E}_P[\cdot]$ to indicate the expectation with respect to measure P .

Linear algebra and matrices

Some chapters depend heavily on linear algebra, though never in much depth. Given $x \in \mathbb{R}^d$ and positive definite matrix $A \in \mathbb{R}^{d \times d}$ the Mahalanobis norm of x with respect to A is $\|x\|_A = \sqrt{x^\top A x}$. The minimum eigenvalue of matrix A is $\lambda_{\min}(A)$, its inverse is A^{-1} , its determinant $\det(A)$, its trace $\text{trace}(A)$ and transpose A^\top . The span of vectors $x_1, \dots, x_n \in \mathbb{R}^d$ is $\text{span}(x_1, \dots, x_n) = \{\sum_i \alpha_i x_i : \alpha \in \mathbb{R}^n\}$. For vectors $x, y \in \mathbb{R}^d$ the inner product is $x^\top y = \langle x, y \rangle = \sum_{i=1}^d x_i y_i$ and the 2-norm is $\|x\|_2 = \langle x, x \rangle^{1/2}$. A matrix $A \in \mathbb{R}^{n \times m}$ is often viewed as a linear map from $\mathbb{R}^m \rightarrow \mathbb{R}^n$. In this case we write $\text{im}(A)$ and $\ker(A)$ for the image and kernel of A respectively. The reader should know Hölder's inequality, which states that $|\langle x, y \rangle| \leq \|x\|_p \|y\|_q$ when $p, q \in [1, \infty]$ are conjugate pairs: $1/p + 1/q = 1$. The most important case is when $p = q = 2$, which yields Cauchy-Schwartz inequality: $|\langle x, y \rangle| \leq \|x\|_2 \|y\|_2$. Unless otherwise mentioned $\|x\|$ is used to denote the 2-norm of x . Jensen's inequality is also indispensable, which says that $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$ for convex functions f . When f is concave the inequality is reversed.

Convexity

A short chapter is devoted to the bare necessities of convexity (Chapter 26). The convex hull of a set is $A \subset \mathbb{R}^d$ is $\text{co}(A)$. For convex set $A \subset \mathbb{R}^d$ and convex function $f : A \rightarrow \mathbb{R}$ we let f^* denote its convex conjugate, $f^*(y) = \sup_{x \in A} \langle x, y \rangle - f(x)$. The support function of A is $\phi_A(x) = \sup_{y \in A} \langle x, y \rangle$ and the polar is the convex set $A^\circ = \{x \in \mathbb{R}^d : \sup_{y \in A} |\langle x, y \rangle| \leq 1\}$.

Landau notation

We make frequent use of Bachmann–Landau notation. Both were 19th century mathematicians who could have never expected their notation to be adopted so enthusiastically by computer scientists. Given functions $f, g : \mathbb{N} \rightarrow [0, \infty)$ define

$$\begin{aligned} f(n) = O(g(n)) &\Leftrightarrow \limsup_{n \rightarrow \infty} \frac{f(n)}{g(n)} < \infty. \\ f(n) = o(g(n)) &\Leftrightarrow \lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0. \\ f(n) = \Omega(g(n)) &\Leftrightarrow \liminf_{n \rightarrow \infty} \frac{f(n)}{g(n)} > 0. \\ f(n) = \omega(g(n)) &\Leftrightarrow \liminf_{n \rightarrow \infty} \frac{f(n)}{g(n)} = \infty. \\ f(n) = \Theta(g(n)) &\Leftrightarrow f(n) = O(g(n)) \text{ and } f(n) = \Omega(g(n)). \end{aligned}$$

We make use of 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 $\lim_{n \rightarrow \infty} f(n)/\sqrt{n} = 1$. 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 $c > 0$ 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.

Sets

\mathbb{N}, \mathbb{N}^+	natural numbers, $\mathbb{N} = \{0, 1, 2, \dots\}$ and $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$
\mathbb{R}	real numbers
$\bar{\mathbb{R}}$	$\mathbb{R} \cup \{-\infty, \infty\}$
2^A	the powerset of set A (the set of all subsets of A)
S_{++}^d	symmetric positive definite matrices in $\mathbb{R}^{d \times d}$
$[n]$	$\{1, 2, 3, \dots, n - 1, n\}$
$\mathcal{P}(A)$	set of probability distributions on discrete set A
A^*	set of finite sequences over A , $A^* = \bigcup_{i=0}^{\infty} A^i$
$\mathfrak{B}(\mathbb{R}^n)$	Borel measurable sets on \mathbb{R}^n
$\mathfrak{L}(\mathbb{R}^n)$	Lebesgue measurable sets on \mathbb{R}^n
B_2^d	d -dimensional unit ball, $\{x \in \mathbb{R}^d : \ x\ _2 \leq 1\}$

Operators

$ A $	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
$\lfloor x \rfloor$	largest integer smaller or equal to x
$\lceil x \rceil$	smallest integer larger or equal to x
$\text{dom}(f)$	domain of function f
\mathbb{E}	expectation
\mathbb{V}	variance
Supp	support function of distribution or random variable
$\nabla f(x)$	gradient of f at x
$\nabla^2 f(x)$	Hessian of f at x
\vee, \wedge	maximum and minimum: $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$

Matrix operations

$\lambda_{\min}(G)$	minimum eigenvalue of matrix G
$\ x\ _G^2$	$x^\top G x$ for positive definite $G \in \mathbb{R}^{d \times d}$ and $x \in \mathbb{R}^d$

Functions

$\text{erf}(x)$	$\frac{2}{\sqrt{\pi}} \int_0^x \exp(-y^2) dy$
$\text{erfc}(x)$	$1 - \text{erf}(x)$

$$\Gamma(z) \quad \text{Gamma function: } \Gamma(z) = \int_0^\infty x^{z-1} \exp(-x) dx$$

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 $[a, b]$
$\text{Beta}(\alpha, \beta)$	Beta distribution with parameters $\alpha, \beta > 0$
δ_x	Dirac distribution with point mass at x

Miscellaneous

∂A	boundary of a set A
$\text{cl}(A)$	closure of set A
$\text{int}(A)$	interior of set A
$\text{co}(A)$	convex hull of A
$\text{aff}(A)$	affine hull of A
$\text{ri}(A)$	relative interior of A
A°	polar of A
e_1, \dots, e_d	standard basis vectors of the d -dimensional Euclidean space
$\text{span}(v_1, \dots, v_d)$	span of vectors v_1, \dots, v_d
$\binom{n}{k}$	binomial coefficient
$\mathbf{0}, \mathbf{1}$	vectors whose elements are all zeros and all ones, respectively.

Part I

Bandits, Probability and Concentration

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

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 [Thompson, 1933]. 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 [Bush and Mosteller, 1953]. 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 will 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 payoff with the distribution of payoffs 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).

There are many reasons to care about bandit problems. Decision making with uncertainty is a challenge we all face and bandits are the simplest example of this demon. Bandit problems also have practical applications. We already mentioned clinical trial design, which researchers have used to motivate their work for eighty 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 would include news recommendation, dynamic pricing and ad placement. As of writing of the book, Google analytics even offers running multi-armed bandit based

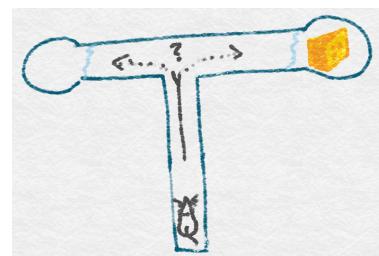


Figure 1.1 Mouse learning a T-maze.

service for their users. A bandit algorithm plays a role in Monte-Carlo Tree Search, an algorithm made famous by the recent success of AlphaGo [Kocsis and Szepesvári, 2006, Silver et al., 2016].

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 information theory, convex analysis/optimization, 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 lead 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 2001–2005, 2006–2010, and 2011–2015 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.

Imagine you are playing a two-armed bandit machine and you already pulled each lever 5 times, resulting in the following payoffs (in dollars):

$$\begin{array}{ll} \text{Left arm:} & 0, 10, 0, 0, 10 \\ \text{Right arm:} & 10, 0, 0, 0, 0 \end{array}$$

The left arm appears to be doing slightly better. The average payoff for this arm is 4 dollars per round, while the average for the right arm is only 2 dollars per round. Let's say, you have 20 more trials (pulls) altogether. How would you pull the arms in the remaining trials? 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 the heart of all bandit problems.



Figure 1.2
Two-armed bandit

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 $n \in \mathbb{N}^+$ is a positive natural number called the **horizon**. In each round the learner first chooses an action A_t from a given set \mathcal{A} and the environment then reveals a reward $X_t \in \mathbb{R}$.

Of course the learner cannot peek into the future when choosing their

actions, which means that A_t should only depend on the **history** $H_{t-1} = (A_1, X_1, \dots, A_{t-1}, X_{t-1})$. A **policy** is a mapping from histories to actions. An environment is a mapping from history sequences ending in actions to rewards. Both the learner and the environment may randomize 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, so 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 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 $\pi \in \Pi$.

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 due to its lack of knowledge of the true environment. 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.

EXAMPLE 1.1 Suppose the action-set is $\mathcal{A} = \{1, 2, \dots, K\}$. 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 characterized 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 policies $\Pi = \{\pi_1, \dots, \pi_K\}$ where π_k chooses action k 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 while the second term is the expected reward collected by the learner.

For a fixed policy and competitor class the regret depends 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 $\lim_{n \rightarrow \infty} R_n/n = 0$. 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 themselves? We will discover eventually that for the environment class in Example 1.1 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 1.1 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 too much. 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. 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 to explain patterns in data and this may be sufficient reason to keep it as the modeling 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 modeling 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 that the learner is not much worse than any of these. By defining the regret in this way we move the stationarity assumption into the definition of regret rather than the environment.

Of course there are all shades of gray between these two extremes. Sometimes we consider the case where the rewards are stochastic, but not stationary. Or one may analyze the robustness of an algorithm for stochastic bandits to small adversarial perturbations. Another idea is to isolate exactly which properties of the stochastic assumption are really exploited by the policies design for stochastic bandits. This kind of inverse analysis can help explain the strong performance of policies when facing environments that clearly violate the assumptions they were designed for.

1.1.1 Why the regret?

One might wonder why bother with the regret at all? If all we really care about is the cumulative rewards, perhaps we should just state our theorems in terms of the rewards. The first observation is that nothing is lost by considering the regret, which simply translates the expected cumulative reward by some environment-dependent constant. There are a few reasons why the regret is useful. One is that it supplies a degree of normalization because it is invariant under translation of rewards. Another benefit is the interpretation as the price paid by the learner for not knowing the true environment. Be warned, however, that this only holds if the competitor class includes the optimal policy.

1.1.2 Other learning objectives

We already mentioned that the regret can be defined in several ways, each capturing slightly different aspects of the behavior of a policy. Because the regret depends on the environment it becomes a multi-objective criteria. One way to convert a multi-objective criteria into a single number is to take averages. This corresponds to the Bayesian viewpoint where the objective is to minimize the average cumulative regret with respect to a prior on the environment class.

Maximizing the sum of rewards is not always the objective. Sometimes the learner just wants to find a near-optimal policy after n rounds, but the actual

rewards accumulated over those rounds are unimportant. We will see examples of this shortly.

1.1.3 Limitations of the bandit framework

The presentation in this section makes it seem like bandits can model almost anything. One of the distinguishing features of all bandit problems studied in this book is that the learner never needs to plan for the future. More precisely, we will invariably make the assumption that the learner's choices and rewards tomorrow are not affected by their decisions today. Problems that require this kind of long-term planning fall into the realm of **reinforcement learning**, which is the topic of the final chapter. Assuming away the need to plan *is* limiting, but as we shall see, it buys you a great deal in terms of simplicity and fits with many applications.

1.2 Applications

After this short preview, and as an appetizer before the hard work, we briefly describe the formalizations of a variety of applications. Remember that for bandit problem we need to choose an action-set and an environment class. Defining the required demands us to provide a competitor class.

A/B testing

The designers of a company website are trying to decide whether the 'buy it now' button should be placed at the top of the product page or the bottom. In the old days they would commit to a trial of each version, where incoming users are split into two groups of ten thousand. Each group is shown a different version of the site and a statistician examines the data at the end to decide which version is better. A problem is the non-adaptivity of the test. Specifically, if the effect size is very large, then the trial could be stopped earlier.

One way to apply bandits to this problem is to view the two versions of the site as actions. Each time t a user makes a request, a bandit algorithm is used to choose an action $A_t \in \mathcal{A} = \{\text{SITEA}, \text{SITEB}\}$ and the reward is $X_t = 1$ if the user purchased the product and $X_t = 0$ otherwise. As opposed to the previous setup, bandits also change the objective: Whereas in A/B testing the goal is to gain knowledge, in the bandit setup the goal is to gain reward. In particular, one should think of running a bandit algorithm continuously as opposed to just running the algorithm until it is able to find out which of the two options is better. In particular, while there are quite a few design choices to be made as we shall see later, life gets simpler as there is no need to decide about a stopping criteria, which in practice is not so simple to decide about ahead of time.

Advert placement

In advert placement each round corresponds to a user visiting a website and the set of actions \mathcal{A} is the set of all available adverts. One could treat this as a standard multi-armed bandit problem, where in each round a policy chooses $A_t \in \mathcal{A}$ and the reward is $X_t = 1$ if the user clicked on the advert and $X_t = 0$ otherwise. This might work for specialized websites where the adverts are all likely to be appropriate. But for a company like Amazon the advertising should be targeted. If I bought rock climbing shoes recently then I'm much more likely to buy a harness than another user. Clearly an algorithm should take this into account.

The standard way to incorporate this additional knowledge is to use the information about the user as **context**. In its simplest formulation this might mean clustering users and implementing a separate bandit algorithm for each cluster. Much of this book is devoted to the question of how to use side information to improve the performance of a learner.

This is a good place to emphasize that the world is messy. The set of available adverts is changing from round to round. The feedback from the user can be delayed for many rounds. Clearly the problem is far from stochastic on large timescales. The adverts that were relevant last year are surely not relevant now. Finally, the metrics to be optimized are rarely as simple as maximizing just for number of clicks (user satisfaction, retention and many other issues matter). These are the kinds of issues that makes implementing bandit algorithms in the real world a difficult task. As noted beforehand, this book will not discuss how to address these issues in detail, but we will rather focus on building up foundations upon which the reader can build and invent new approaches to address the messiness of their own real world problems.

Recommendation services

Netflix has to decide which movies to place most prominently in your ‘Browse’ page. Like in advert placement the users arrive at the page sequentially and the reward can be measured using (a) whether or not you watched a movie and (b) whether or not you rated it positively. There are many challenges. First of all, Netflix doesn’t just recommend one movie, they show you a list. 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 factorization (a popular approach in “collaborative filtering”), but notice that this is not an offline problem. The learning algorithm gets to choose what the users see and this affects the data that is collected. 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 with even relatively small graphs possessing an enormous number of paths. The routing problem can obviously applied to more physical networks such as transportation systems used in operations research.

Dynamic pricing

In dynamic pricing a company is trying to automatically optimize 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 five minutes, but the timetable is unreliable and the bus arrival time unknown and stochastic. Sometimes the bus doesn't come at all. Walking, on the other hand, takes thirty 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. 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 came, which is a problem analyzed 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 problem interesting.

Resource allocation

High speed cache memory is still a scarce resource for computer processors and the consequence in terms of running time for cache misses is quite extreme. Many algorithms are optimized to match the cache process, for example by carefully choosing the order of dimensions in arrays or using cache-oblivious trees. A less-explored avenue of improvement is to try and learn the optimal allocation

of cache resources between processes. This can be modeled by a bandit problem where the set of actions is the set of allocations and the reward is the negation of the number of cache misses. For this problem the learner might reasonably make a monotonicity assumption in the sense that increasing the allocation for one process should decrease the number of cache misses.

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 analyzed theoretically, but more importantly has demonstrated outstanding empirical performance in game playing problems.

1.3 Bibliographic remarks

We already mentioned that the first paper on bandits was by [Thompson \[1933\]](#). Much credit for the popularization 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 was 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 six years 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. There are also three books on sequential design and multi-armed bandits in the Bayesian setting, which we will address only a little. Both are based on relatively old material, but are still useful references for this line of work and are well worth reading [\[Chernoff, 1959, Berry and Fristedt, 1985, Gittins et al., 2011\]](#).

Without trying to be exhaustive, here are a few articles applying bandit algorithms to applications. The papers themselves will contain more useful pointers to the vast literature. [Le et al. \[2014\]](#) applies bandits to wireless monitoring where the problem is challenging due to the large action space.

Lei et al. [2017] design specialized 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 behavioral change. See also the article by Greenewald et al. [2017]. Rafferty et al. [2018] applies Thompson sampling to educational software and notes the tradeoff between knowledge and reward. That bandit algorithms have not been used in clinical trials was explicitly noted by Villar et al. [2015]. Microsoft offers a ‘Decision Service’ that uses bandit algorithms to automate decision-making [Agarwal et al., 2016]. We already mentioned that bandit algorithms are a cornerstone of Monte-Carlo Tree Search [Kocsis and Szepesvári, 2006]. Muller et al. [2017] uses bandits for estimating the H_∞ -gain of linear systems; the problem here is to excite a linear control system by designing clever inputs so that the magnitude of the highest frequency amplification in the input is estimated. Knowing the H_∞ -gain is helpful for assessing the robustness of a control loop.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

2 Foundations of Probability (\dagger)

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 rigor 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 doubtless 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

Probability theory is a latecomer to the party of mathematical study. While the ancient Greeks and Romans certainly gambled, there is no evidence they ever formally analyzed the probabilistic nature of the games they played. But probability does have its origins in the study of games of chance and gambling, with early steps taken in the 16th and 17th centuries by famous mathematicians and physicists such as Niccoló Tartaglia, Gerolamo Cardano, Blaise Pascal, Pierre Fermat, Christian Huygens and Jacob Bernoulli. 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 the newly thrown dice (one or two), I repeat the same, for a total of three rounds (and at most seven dice throws in total). 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?

The fact that the number of dice used is random appears to create a messy

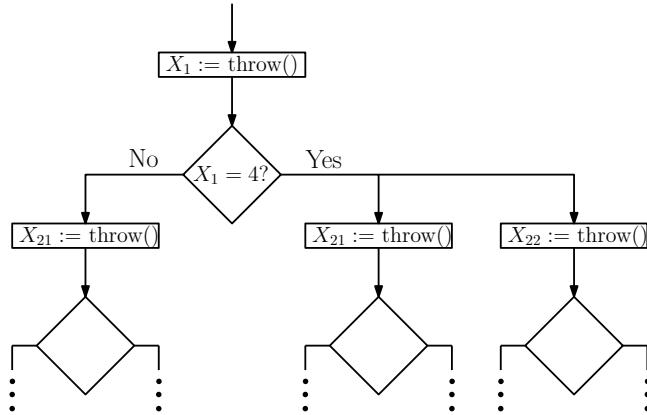


Figure 2.1 The initial phase of a gambling game with a random number of dice rolls. Depending on the outcome of a dice roll, one or two dice are rolled for a total of three rounds. The number of dice used will then be random in the range of three to seven.

situation where the outcomes have a complicated dependency structure. This situation is not unusual, with many examples of practical interest exhibiting the same random interdependency between outcomes. The fundamental idea in modern probability is aimed at removing this complication.

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 (see Fig. 2.1). With the dice all rolled, the game can be emulated easily 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.



This approach separates the randomness (rolls of the dice) from the mechanism that produces values based on the random outcomes. This idea is one of the cornerstones of modern probability as proposed by Kolmogorov.

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 payoff, we first consider the probability of any single outcome of the dice. Since there are seven dice, the set of all possible outcomes is $\Omega = \{1, \dots, 6\}^7$. Because all outcomes are equally probable the probability of any $\omega \in \Omega$ is $(1/6)^7$. The probability of the game payoff 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

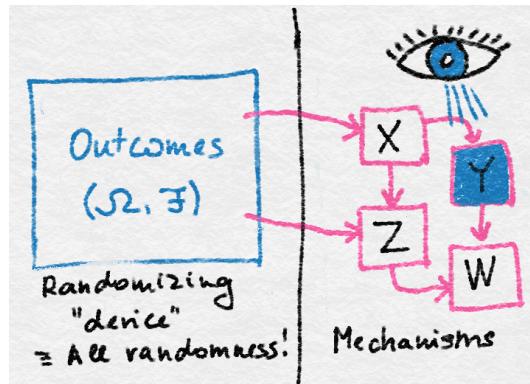


Figure 2.2 A key idea in probability theory is the separation of sources of randomness from game mechanisms. A mechanism creates values from the elementary random outcomes, some of which are visible for observers, while others may remain hidden.

generated on the left, while on the right, various mechanisms are used to arrive at values, some of which 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 value is a function X that maps Ω to the set of natural numbers \mathbb{N} (simply, $X : \Omega \rightarrow \mathbb{N}$). While we view the value of the game as random, this map is deterministic. We find it ironic that functions of this type (from the outcome space to subsets of the reals) are called **random variables**. They are neither random nor variables in a programming language sense. The randomness is in the argument that X is acting on, producing randomly changing results. Later we will put a little more structure on random variables, but for now it suffices to think of them as maps from the outcome space to the reals.



We will follow the standard convention in probability theory where random variables are denoted by capital letters. Be warned that capital letters are also used for other purposes as demanded by different conventions.

Pick some natural number $v \in \mathbb{N}$. What is the probability of seeing $X = v$? As described above, this probability is $(1/6)^7$ times the size of the set $X^{-1}(v) = \{\omega \in \Omega : X(\omega) = v\}$. The set $X^{-1}(v)$ is called the **preimage** of v under X . More generally, the probability that X takes its value in some set $A \subseteq \mathbb{N}$ is given by $(1/6)^7$ times the cardinality of $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$, where we have overloaded the definition of X^{-1} to set-valued inputs.

Notice in the previous paragraph we only needed probabilities assigned to subsets of Ω regardless of the question asked. To make this a bit more general, let us introduce a map \mathbb{P} that assigns probabilities to certain subsets of Ω . The intuitive meaning of \mathbb{P} is as follows. Random outcomes are generated in Ω . The

probability that an outcome falls into a set $A \subset \Omega$ is $\mathbb{P}(A)$. If A is not in the domain of \mathbb{P} , then there is no answer to the question of the probability of the outcome falling in A . But let's postpone the discussion of why \mathbb{P} should be restricted to only certain subsets of Ω later. In the above example with the dice, for any subset $A \subseteq \Omega$, $\mathbb{P}(A) = (1/6)^7 |A|$.

With this new notation, the answer to the question of what is the probability of seeing X taking the value of v becomes $\mathbb{P}(X^{-1}(v))$. To minimize clutter, the more readable notation for this is $\mathbb{P}(X = v)$. It is important to realize, however, that this familiar form is a shorthand for $\mathbb{P}(X^{-1}(v))$. More generally, we also use

$$\mathbb{P}(\text{predicate}(X, U, V, \dots)) = \mathbb{P}(\{\omega \in \Omega : \text{predicate}(X, U, V, \dots) \text{ is true}\})$$

with any predicate (an expression evaluating to true or false) where X, U, V, \dots are functions with domain Ω .

What properties should \mathbb{P} satisfy? It seems reasonable to expect the probability that *something* happens is one, which is equivalent to saying that \mathbb{P} is defined for Ω and $\mathbb{P}(\Omega) = 1$. Second, probabilities should be nonnegative so $\mathbb{P}(A) \geq 0$ for any $A \subset \Omega$ on which \mathbb{P} is defined. Let $A^c = \Omega \setminus A$ be the complement of A . Then we should expect that $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ (negation rule). Finally, if A, B are disjoint so that $A \cap B = \emptyset$ and $\mathbb{P}(A)$, $\mathbb{P}(B)$ and $\mathbb{P}(A \cup B)$ are all defined, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. This is called the **finite additivity property**.

Let \mathcal{F} be the set of subsets of Ω on which \mathbb{P} is defined. It would seem silly if $A \in \mathcal{F}$ and $A^c \notin \mathcal{F}$, since $\mathbb{P}(A^c)$ could simply be defined by $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$. Similarly, if \mathbb{P} is defined on disjoint sets A and B , then it makes sense $A \cup B \in \mathcal{F}$. By a logical jump, we will also require the additivity property to hold for **countably infinitely many** sets. If $\{A_i\}_i$ is a collection of sets and $A_i \in \mathcal{F}$ for all $i \in \mathbb{N}$, then $\cup_i A_i \in \mathcal{F}$ and $\mathbb{P}(\cup_i A_i) = \sum_i \mathbb{P}(A_i)$. A set of subsets that satisfies all these properties is called a **σ -algebra**, which is pronounced ‘sigma-algebra’ and sometimes also called a σ -field.

DEFINITION 2.1 A set $\mathcal{F} \subseteq 2^\Omega$ is a σ -algebra if $\Omega \in \mathcal{F}$ and $A^c \in \mathcal{F}$ for all $A \in \mathcal{F}$ and $\cup_i A_i \in \mathcal{F}$ for all $\{A_i\}_i$ with $A_i \in \mathcal{F}$ for all $i \in \mathbb{N}$. A function $\mathbb{P} : \mathcal{F} \rightarrow \mathbb{R}$ is a **probability measure** if $\mathbb{P}(\Omega) = 1$ and for all $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$ and $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ and $\mathbb{P}(\cup_i A_i) = \sum_i \mathbb{P}(A_i)$ for all countable collections of disjoint sets $\{A_i\}_i$ with $A_i \in \mathcal{F}$ for all i . If \mathcal{F} is a σ -algebra and $\mathcal{G} \subset \mathcal{F}$ is also a σ -algebra, then we say \mathcal{G} is a **sub- σ -algebra** of \mathcal{F} .

The elements of \mathcal{F} are called **measurable sets**. They are measurable in the sense that \mathbb{P} assigns values to them. The pair (Ω, \mathcal{F}) alone is called a **measurable space**, while the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**. If the condition that $\mathbb{P}(\Omega) = 1$ is lifted, then \mathbb{P} is called a **measure**. If the condition that $\mathbb{P}(A) \geq 0$ is also lifted, then \mathbb{P} is called a **signed measure**. We note in passing that both for measures and signed measures it would be unusual to use the symbol \mathbb{P} , which is mostly reserved for probabilities.

Random variables lead to new probability measures. In particular, $\mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A))$ is a probability measure defined for all the subsets A of \mathbb{N} for which

$\mathbb{P}(X^{-1}(A))$ is defined. The probability measure \mathbb{P}_X is called the **probability measure induced by X** and \mathbb{P} , or the **pushforward** measure of \mathbb{P} under X . An important observation is that any probabilistic question concerning X can be answered from the knowledge of \mathbb{P}_X alone.

It is worth noting that if we keep X fixed, but change \mathbb{P} (for example, by switching to loaded dice), then the measure induced by X changes. We will often use arguments that do exactly this, especially when proving lower bounds on the limits of how well bandit algorithms can perform.

The astute reader would have noticed that we skipped over some details. In particular, we defined measures as functions from a σ -algebra to \mathbb{R} . So if we want to call \mathbb{P}_X a measure, then its domain $\{A \subset \mathbb{N} : X^{-1}(A) \in \mathcal{F}\}$ better be a σ -algebra. This is indeed the case and in fact it holds even more generally that sets of the above form are σ -algebras regardless of the range of X (Exercise 2.1).

Let (Ω, \mathcal{F}) be a measurable space, \mathcal{X} be an arbitrary set and $\mathcal{G} \subseteq 2^{\mathcal{X}}$. A function $X : \Omega \rightarrow \mathcal{X}$ is called a **\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{G} is 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} 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 2.3). 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**. More generally, the Borel σ -algebra on a topological space $(\mathcal{X}, \mathcal{U})$ is the σ -algebra generated by its open sets $\sigma(\mathcal{U})$.

DEFINITION 2.2 A **random variable** on measurable space (Ω, \mathcal{F}) is a $\mathcal{F}/\mathfrak{B}(\mathbb{R})$ -measurable function $X : \Omega \rightarrow \mathbb{R}$. A **random element** between measurable spaces (Ω, \mathcal{F}) and $(\mathcal{X}, \mathcal{G})$ is a \mathcal{F}/\mathcal{G} -measurable function $X : \Omega \rightarrow \mathcal{X}$.

The only difference between random variables and elements is that the former are restricted to Borel-measurable functions taking values in the reals. A **Borel function** is any function $f : \mathbb{R} \rightarrow \mathbb{R}$ that is $\mathfrak{B}/\mathfrak{B}$ -measurable.

Indicator functions

In the following text we make heavy uses of **indicator functions**. Given an arbitrary set Ω and $A \subseteq \Omega$ the indicator function of A is $\mathbb{I}_A : \Omega \rightarrow \{0, 1\}$ given by

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

To abuse notation even further but with the noble goal of removing clutter, we will often write $\mathbb{I}\{\omega \in A\}$ or $\mathbb{I}\{\text{predicate}(X, Y, \dots)\}$, with the latter being the indicator function of the set on which the predicate is true. It is easy to check that an indicator function \mathbb{I}_A is a random variable on (Ω, \mathcal{F}) if and only if $A \in \mathcal{F}$.

Why so complicated?

You may be wondering why we did not define \mathbb{P} on the powerset of Ω , which is equivalent to declaring all sets to be measurable. In many cases this is a perfectly reasonable thing to do, including the example game above where nothing prevents us from defining $\mathcal{F} = 2^\Omega$. There are two justifications not to do this, the first technical and the second conceptual. The technical issue is highlighted by the following surprising theorem, which shows there does not exist a uniform probability distribution on $\Omega = [0, 1]$ if \mathcal{F} is chosen to be the powerset of Ω . 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 on the Borel σ -algebra, though it is not as easy as you might expect.

THEOREM 2.1 *Let $\Omega = [0, 1]$ and \mathcal{F} is the powerset of Ω . Then there does not exist a measure \mathbb{P} on (Ω, \mathcal{F}) such that $\mathbb{P}([a, b]) = b - a$ for all $0 \leq a \leq b \leq 1$.*

There are other technical reasons besides this. For example, measure theory allows for the unification of distributions on discrete spaces and densities on continuous ones. This can be necessary when one wants to talk about measures that combine elements of both (for example, draw a random number of a normal whose mean is chosen at random from a finite set of means). The main conceptual reason not to focus exclusively on the case where \mathcal{F} is the powerset is that σ -algebras are a way of representing information. This is especially useful in the study of bandits where the learner is interacting with an environment and slowly gaining knowledge. One useful way to represent this idea is by the means of a sequence of σ -algebras as we explain in the next section. One might be worried that the Borel σ -algebra does not contain enough measurable sets. Rest assured that this is not a problem and you will be hard pressed to find a non-measurable set. An example is given in the notes, along with a little more discussion on this topic.

Probability spaces from random variables

The 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

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

which represents the joint distribution for the values of a dice ($X \in [6]$) and coin ($Y \in [2]$). The formula describes the probabilistic interactions between the outputs of X and Y , but says nothing about their domain. The follow theorem gives conditions under which a probability space carrying X and Y exists.

THEOREM 2.2 *Let $n \in \mathbb{N}^+$ and for each $k \in [n]$ let $(\Omega_k, \mathcal{F}_k)$ be a measurable space and let $\bar{\mu} : \mathcal{F}_1 \times \cdots \times \mathcal{F}_n \rightarrow [0, 1]$ be a function such that:*

- (a) $\bar{\mu}(\Omega_1 \times \cdots \times \Omega_n) = 1$.
 (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 (Ω, \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.



Even when $n = 2$ it is not true that $\mathcal{F}_1 \times \mathcal{F}_2 = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ since the former set does not contain $\{(1, 1), (2, 2)\}$.

The **law** of a random variable X on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ given by \mathbb{P}_X with $\mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A))$. More generally, when $(\mathcal{X}, \mathcal{G})$ is a measurable space and $X : \Omega \rightarrow \mathcal{X}$ is a \mathcal{F}/\mathcal{G} -measurable random element, then the law of X is a measure on $(\mathcal{X}, \mathcal{G})$. Finally, when $(X_k)_{k=1}^n$ is a collection of random elements to measurable spaces $(\mathcal{X}_k, \mathcal{G}_k)$, then the law of $(X_k)_{k=1}^n$ is a measure on $(\mathcal{X}_1 \times \cdots \times \mathcal{X}_n, \sigma(\mathcal{G}_1 \times \cdots \times \mathcal{G}_n))$.

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’. Regrettably this relationship is not a perfect one, but nevertheless it is useful and provides a great deal of intuition. Let (Ω, \mathcal{F}) , $(\mathcal{X}, \mathcal{G})$ and $(\mathcal{Y}, \mathcal{H})$ be measurable spaces and $X : \Omega \rightarrow \mathcal{X}$ and $Y : \Omega \rightarrow \mathcal{Y}$ be random elements. The σ -algebra generated by X is defined by

$$\sigma(X) = \{X^{-1}(A) : A \in \mathcal{G}\} \subseteq \mathcal{F}.$$

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. Having observed the value of 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 ? Except for some technical assumptions 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.

LEMMA 2.1 (Factorization lemma) *Assume that $(\mathcal{Y}, \mathcal{H})$ is Borel. Then Y is $\sigma(X)$ -measurable if and only if there exists a \mathcal{G}/\mathcal{H} -measurable map $f : \mathcal{X} \rightarrow \mathcal{Y}$ such that $Y = f \circ X$.*

$$\begin{array}{ccc} (\Omega, \mathcal{F}) & \xrightarrow{X} & (\mathcal{X}, \mathcal{G}) \\ & \searrow Y & \downarrow f \\ & & (\mathcal{Y}, \mathcal{H}) \end{array}$$

What this means is that having observed X the value of Y can be computed as $f(X)$ for some measurable function f if and only if Y is $\sigma(X)$ -measurable. In this sense $\sigma(X)$ contains all the information that can be extracted from X via measurable functions. This is not quite the same as saying that Y can be deduced from X if and only if Y is $\sigma(X)$ -measurable, because the definition prohibits the use of nonmeasurable functions. In extreme cases the restriction to measurable functions can be a severe limitation.

EXAMPLE 2.1 Let $\Omega = \mathcal{Y} = \mathcal{X} = \mathbb{R}$, $\mathcal{F} = \mathcal{H} = \mathfrak{B}(\mathbb{R})$ and $\mathcal{G} = \{\emptyset, \mathbb{R}\}$ be the trivial σ -algebra. In this case \mathcal{G} is so sparse that all functions $X : \Omega \rightarrow \mathcal{X}$ are \mathcal{F}/\mathcal{G} -measurable and we will choose $X(\omega) = Y(\omega) = \omega$ to be the identity functions. Now $\sigma(X) = \{\emptyset, \mathbb{R}\}$ and Y is not $\sigma(X)/\mathfrak{B}(\mathbb{R})$ -measurable and the lemma says there does not exist a $\mathcal{G}/\mathfrak{B}(\mathbb{R})$ -measurable function f with $Y = f \circ X$. But here $Y = X$ and the knowledge of Y is easily deduced from X . The problem is that the process by which this deduction takes place is not measurable because \mathcal{G} is not sufficiently rich.

The above example emphasizes the point 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 course-grained, then $\sigma(X)$ can also be course 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 $\mathcal{F} = \mathcal{G} = \mathcal{H} = \mathfrak{B}(\mathbb{R})$ it can still occur that $Y = f \circ X$ for some nonmeasurable 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 nonmeasurable sets in \mathcal{X} . While such random variables exist, they are almost never encountered in practice and when a probability measure \mathbb{P} is introduced one can invariably prove the existence of a measurable \bar{f} such that $\mathbb{P}(Y = \bar{f} \circ X) = 1$.

Filtrations

In the study of bandits and other online settings it usually occurs that information is revealed to the learner sequentially. Let X_1, \dots, X_n be a collection of random variables on common measurable space (Ω, \mathcal{F}) . We imagine a learner is sequentially observing the values of these random variables. First X_1 , then X_2 and so on. Having observed X_1 the learner has access to the information in $\sigma(X_1)$. By this we mean that the value of $\sigma(X_1)$ -measurable maps can be deduced from the value of X_1 using a measurable function. Then the learner observes X_2 and now they have access to $\sigma(X_1, X_2)$. This latter quantity is the smallest σ -algebra for which

both X_1 and X_2 are measurable. Generalizing further, let $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ be the σ -algebra containing information observed after time t . It is easy to check that $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots \subseteq \mathcal{F}_n$, which means that more and more functions are becoming \mathcal{F}_t -measurable as t increases, which corresponds to increasing knowledge.

Often we want to talk about increasing sequences of σ -algebras without constructing them in terms of random variables as above. Given measurable space (Ω, \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 $t < n$. Note that 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 in the obvious way, 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. Finally, $(X_t)_t$ is \mathbb{F} -**predictable** if X_t is \mathcal{F}_{t-1} -measurable for each $t \in [n]$. Intuitively (with the caveats expressed earlier), we may think of a \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. For the formal definition, let $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the space of outcomes, \mathcal{F} is the collection of events to which the probability measure \mathbb{P} assigns probabilities. Fix some event $B \in \mathcal{F}$ that has a positive probability and consider some other event $A \in \mathcal{F}$. The **conditional probability** of A given B , and denoted by $\mathbb{P}(A | B)$, is defined as

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

This answers the question how the knowledge that B occurred changes the probability assigned to A . We can think about the outcome $\omega \in \Omega$ as the result of throwing a many-sided dice. The question asked is the probability that the dice landed so that $\omega \in A$ given that it landed with $\omega \in B$. The meaning of the condition $\omega \in B$ is that we focus on dice rolls when $\omega \in B$ is true. All dice rolls when $\omega \in B$ does not hold are discarded. Intuitively, what should matter in the

conditional probability of A given B is how large the portion of A is that lies in B and this is indeed what the definition means.



The importance of conditional probabilities is that they define a calculus of how probabilities are to be updated in the presence of extra information.

To emphasize this relationship to knowledge, the probability $\mathbb{P}(A | B)$ is also called the **a posteriori** ('after the fact') probability of A given B . In contrast, its **a priori** probability is $\mathbb{P}(A)$. Note that $\mathbb{P}(A | B)$ is defined for every $A \in \mathcal{F}$ as long as $\mathbb{P}(B) > 0$. In fact, $A \mapsto \mathbb{P}(A | B)$ is a probability measure over the measure space (Ω, \mathcal{F}) called the a posteriori probability measure given B (see Exercise 2.4). In a way the temporal characteristics attached to the words 'a posteriori' and 'a priori' can be a bit misleading. As discussed before, probabilities are concerned with predictions. They express the degrees of uncertainty one assigns to future events. This is also true for conditional probabilities. The conditional probability of A given B is a prediction of certain properties of the outcome of the random experiment that results in ω given a certain condition. Note that everything is related to a future hypothetical outcome. Once the dice is rolled, ω gets fixed and either $\omega \in B$ or not, and either $\omega \in A$ or not. There is no randomness, no uncertainty left. Probability theory is thus a science of predictions – this is where its power is coming from (but one cannot talk about predictions after an experiment is done).

We carefully avoided the issue of how to define conditional probabilities when an event has zero probability. While it may seem weird at a first sight to even ponder about the proper definition for this case, it turns out that conditional probabilities that are also defined (in some yet unspecified way) have some useful properties which makes the effort of defining them more than worthwhile. The discussion of this, however, has to wait until after we have introduced conditional expectations.

The discussion of conditional probabilities would not be complete without mentioning **Bayes law** or **Bayes rule**, which states that provided $A, B \in \mathcal{F}$ have both positive probabilities,

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(B | A) \mathbb{P}(A)}{\mathbb{P}(B)}. \quad (2.1)$$

This is obviously useful when we want to know $\mathbb{P}(A | B)$ and we have information about the quantities on the right-hand side. Remarkably, this happens to be the case quite often, explaining why this simple formula has quite a status in probability and statistics. Exercise 2.5 asks the reader to verify this law.

2.4 Independence

Independence is another basic concept of probability that relates to knowledge/information. The easiest way to define independence is through conditional probabilities. In its simplest form independence is a relation that holds between events on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Two events $A, B \in \mathcal{F}$ are **independent** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \quad (2.2)$$

How is this related to knowledge? Assuming that $\mathbb{P}(B) > 0$, dividing both sides by $\mathbb{P}(B)$ and using the definition of conditional probability we get that the above is equivalent to

$$\mathbb{P}(A | B) = \mathbb{P}(A). \quad (2.3)$$

Of course, we also have that if $\mathbb{P}(A) > 0$, (2.2) is equivalent to $\mathbb{P}(B | A) = \mathbb{P}(B)$. Note that both of the latter relations express something intuitive, which is that A and B are independent if the probability assigned to A (or B) remains the same regardless of whether it is known that B (respectively, A) occurred.



Independence of two events means that observing the outcome of one does not change the likelihood of the other.

We hope our readers will find the definition of independence in terms of a ‘lack of influence’ to be sensible. The reason not to use Eq. (2.3) as the definition is mostly for the sake of convenience. If we started with (2.3) we would need to separately discuss the case of $\mathbb{P}(B) = 0$, which would be cumbersome. A second reason is that (2.3) suggests an asymmetric relationship, but intuitively we expect independence to be symmetric.

Why care about independence? There are at least two, unrelated reasons: (i) Uncertain outcomes are often generated part by part with no interaction between the processes, which naturally leads to an independence structure (think of rolling multiple dice with no interactions between the rolls) and (ii) once we discover some independence structure, calculations with probabilities can be immensely simplified. In fact, independence is often used as a way of constructing probability measures of interest (Exercise 2.6). Independence can also appear serendipitously in the sense that a probability space may hold many more independent events than what its construction may obviously suggest (Exercise 2.7).

Independence assumptions should not be taken lightly. Whenever independence is brought up, one should carefully judge whether the independence structure is really true. Since this is part of modeling, this reasoning is not mathematical in nature but is concerned with thinking about the physical processes.

A collection of events $\mathcal{G} \subset \mathcal{F}$ is said to be **pairwise independent** if any two distinct elements of \mathcal{G} are independent of each other. The events in \mathcal{G} are said

to be **mutually independent** if for any $n > 0$ integer and A_1, \dots, A_n distinct elements of \mathcal{G} , $\mathbb{P}(A_1 \cap \dots \cap A_n) = \prod_{i=1}^n \mathbb{P}(A_i)$. This is a stronger restriction than pairwise independence. In the case of mutually independent events the knowledge of joint occurrence of any finitely many events from the collection will not change our prediction of whether some other event happens. But this may not be the case when the events are only pairwise independent (Exercise 2.7). Two **collections of events** $\mathcal{G}_1, \mathcal{G}_2$ are said to be **independent of each other** if for any $A \in \mathcal{G}_1$ and $B \in \mathcal{G}_2$ it holds that A and B are independent of each other. This definition is often applied to σ -algebras.

When the σ algebras are induced by random variables, this leads to the definition of **independence between random variables**. Two random variables X and Y are called independent of each other if their underlying σ -algebras, $\sigma(X)$ and $\sigma(Y)$, are independent of each other. As we discussed previously, σ -algebras summarize what knowledge can be gained by learning the value of a random variable. Hence the above definition says that two random variables are independent of each other when learning the value of one of them does not help in any way predicting the value of the other. The notions of pairwise and mutual independence can also be naturally extended to apply to collections of random variables. All these concepts can be and are in fact extended to random elements. The default meaning of independence when multiple events or random variables are involved is mutual independence. Thus, when we say that X_1, \dots, X_n are independent random variables, we mean that they are mutually independent. When discussing independence, the probability measure is often hidden.



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**, or **mean value** of random variables. For the formal definition fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The expectation of a random variable $X : \Omega \rightarrow \mathbb{R}$ 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} .

$$\mathbb{E}[X] = \int X(\omega) d\mathbb{P}(\omega).$$

The right-hand side is also often abbreviated to $\int X d\mathbb{P}$, suppressing the variable ω that the integration is over. Shortly we will construct the integral on the right hand side to satisfy the following 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 $\int X d\mathbb{P} = \mathbb{P}(A)$.
- (b) Integrals are linear. For all random variables X_1, X_2 and reals α_1, α_2 such that $\int X_1 d\mathbb{P}$ and $\int X_2 d\mathbb{P}$ are defined, $\int(\alpha_1 X_1 + \alpha_2 X_2) d\mathbb{P}$ is defined and satisfies

$$\int(\alpha_1 X_1 + \alpha_2 X_2) d\mathbb{P} = \alpha_1 \int X_1 d\mathbb{P} + \alpha_2 \int X_2 d\mathbb{P}. \quad (2.4)$$

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 $\int X d\mathbb{P} = \sum_i \alpha_i \mathbb{P}(A_i)$. Functions of this type are called **simple functions**.

The next step is to extend the definition to nonnegative random variables. Let $X : \Omega \rightarrow [0, \infty)$ be measurable. The idea is to approximate X using simple functions from below and take the largest value that can be obtained this way.

$$\int_{\Omega} X d\mathbb{P} = \sup \left\{ \int_{\Omega} h d\mathbb{P} : h \text{ is simple and } 0 \leq h \leq X \right\}. \quad (2.5)$$

Here $h \leq X$ if $h(\omega) \leq X(\omega)$ 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 $X : \Omega \rightarrow \mathbb{R}$ 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 nonnegative random variables called the **positive** and **negative** parts of X . Provided that both X^+ and X^- are integrable we define

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

Note that X is integrable if and only if the nonnegative value random variable $|X|$ is integrable. The reader may challenge themselves by proving this in Exercise 2.9.

None of what we have done depends on \mathbb{P} being a probability measure (that is $\mathbb{P}(A) \geq 0$ and $\mathbb{P}(\Omega) = 1$). The definitions all hold more generally 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} is the so-called Lebesgue σ -algebra (defined in the notes below), while the measure is the so-called Lebesgue measure λ , which is the unique measure such that $\lambda((a, b)) = b - a$ for any $a \leq b$. In this scenario, if $f : \mathbb{R} \rightarrow \mathbb{R}$ is a

Borel-measurable function (again, for the definition, see below), 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 do, however, exist functions that are Riemann integrable and not Lebesgue integrable, and also the other way around (although examples of the former are more unusual than the latter). This 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 is their linearity, which was postulated above as the second property in (2.4). 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 2.1 *Let $(X_i)_i$ be a (possibly infinite) collection 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. We will not prove this statement here, but as usual suggest the reader do so for themselves (Exercise 2.11). The other requirement for linearity is that if $c \in \mathbb{R}$ is a constant, then $\mathbb{E}[cX] = c\mathbb{E}[X]$, which is also true and rather easy to prove (Exercise 2.12). Another important statement is concerned with independent random variables.

PROPOSITION 2.2 *If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

Note that in general $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$ (Exercise 2.14). Finally, an important simple result connects expectations of nonnegative random variables to their tail probabilities.

PROPOSITION 2.3 *If $X \geq 0$ is a nonnegative random variable, then*

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

The integrand in Proposition 2.3 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 nonnegative ones. One can check that $F_X : \mathbb{R} \rightarrow [0, 1]$ is nondecreasing, right-continuous and $\lim_{x \rightarrow -\infty} F_X(x) = 0$ and $\lim_{x \rightarrow \infty} F_X(x) = 1$. 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 .

2.6 Conditional expectation

Besides the expectation, we will also need **conditional expectation**, which allows us to talk about the expectation of a random variable given the value of another random variable. To illustrate with an example, 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 $\mathbb{E}[X | Y]$. Using this notation, in the above example, we can concisely write $\mathbb{E}[X | Y] = 5Y + 2(1 - Y)$. A little more generally, if $X : \Omega \rightarrow \mathcal{X}$ and $Y : \Omega \rightarrow \mathcal{Y}$ with $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$ and $|\mathcal{X}|, |\mathcal{Y}| < \infty$. Then $\mathbb{E}[X | Y] : \Omega \rightarrow \mathbb{R}$ is the random variable given by $\mathbb{E}[X | Y](\omega) = \mathbb{E}[X, Y = Y(\omega)]$ where

$$\mathbb{E}[X | Y = y] = \sum_{x \in \mathcal{X}} x \mathbb{P}(X = x | Y = y) = \sum_{x \in \mathcal{X}} \frac{x \mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}. \quad (2.6)$$

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

The definition in Eq. (2.6) does not generalize to continuous random variables because $\mathbb{P}(Y = y)$ 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 $\mathbb{P}(X = 1, Y \in [p, q]) = \int_p^q x dx$ for $0 \leq p < q \leq 1$. Intuitively it seems like $\mathbb{E}[X | Y]$ should be equal to Y , but how to define it? Remember that the mean of a Bernoulli random variable is equal to its bias so the definition of

conditional probability shows that for $0 \leq p < q \leq 1$,

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

The above calculation is not well defined when $p = q$ because $\mathbb{P}(Y \in [p, p]) = 0$. 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 $\mathbb{P}(X = 1 \mid Y = p) = p$. Unfortunately this approach does not generalize 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.

From Eq. (2.6) 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 $A \subseteq \mathcal{Y}$ the above display shows that

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

In fact these two properties alone completely determine $\mathbb{E}[X \mid Y]$ except for a set of measure zero. Notice that both conditions actually only depend on $\sigma(Y) \subseteq \mathcal{F}$. Summarizing, we expect that $\mathbb{E}[X \mid Y]$ be $\sigma(Y)$ -measurable and for all $B \in \sigma(Y)$, $\mathbb{E}[\mathbb{I}_B \mathbb{E}[X \mid Y]] = \mathbb{E}[\mathbb{I}_B X]$. The abstract definition of conditional expectation takes these properties as the definition and replaces the role of Y with a sub- σ -algebra.

DEFINITION 2.3 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \rightarrow \mathbb{R}$ 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 $H \in \mathcal{H}$,

$$\int_H \mathbb{E}[X \mid \mathcal{H}] d\mathbb{P} = \int_H X d\mathbb{P}. \quad (2.7)$$

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



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 2.1 and the fact that $\mathbb{E}[X \mid \sigma(Y)]$ is $\sigma(Y)$ -measurable shows there exists a measurable function $f : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$\mathbb{E}[X \mid \sigma(Y)](\omega) = (f \circ Y)(\omega)$ (see Fig. 2.3). In this sense $\mathbb{E}[X \mid Y](\omega)$ only depends on $Y(\omega)$ and occasionally we write $\mathbb{E}[X \mid Y](y)$.

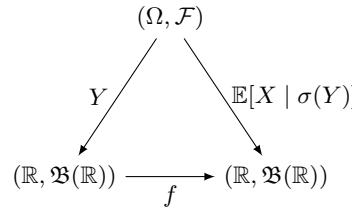


Figure 2.3 Factorization of conditional expectation.

At the risk of being overly verbose, what is the meaning of all this? Returning to the dice example above we see that $\mathbb{E}[X \mid Y] = \mathbb{E}[X \mid \sigma(Y)]$ and $\sigma(Y) = \{\{1, 2, 3\}, \{4, 5, 6\}, \emptyset, \Omega\}$. The condition that $\mathbb{E}[X \mid \mathcal{H}]$ is \mathcal{H} -measurable can only be satisfied if $\mathbb{E}[X \mid \mathcal{H}](\omega)$ is constant on $\{1, 2, 3\}$ and $\{4, 5, 6\}$. Then (2.7) immediately implies that

$$\mathbb{E}[X \mid \mathcal{H}](\omega) = \begin{cases} 2, & \text{if } \omega \in \{1, 2, 3\}; \\ 5, & \text{if } \omega \in \{4, 5, 6\}. \end{cases}$$

Finally, we want to emphasize that the definition of conditional expectation given above is not constructive. Even more off-putting is that $\mathbb{E}[X \mid \mathcal{H}]$ is not even uniquely defined, though any two conditional expectations will only differ in a set of measure zero, which does not matter when we calculate further expectations.

A related notation that will be useful in the future is the concept of **almost surely**. Let X and Y be two random variables on the same probability space. Then $X = Y$ \mathbb{P} -almost surely if $\mathbb{P}(X = Y) = 1$. This is often abbreviated to $X = Y$ \mathbb{P} -a.s. or just $X = Y$ a.s. when the probability measure is clear from the context. Yet another alternative is to write $X = Y$ with **probability one** or ‘with \mathbb{P} -probability one’. Note that this agrees with $\mathbb{P}(X \neq Y) = 0$.

In the above examples we were ‘lucky’ to find the random variable that satisfies the definition of conditional expectation. Or were we? In other words, do conditional expectations always exist? Unsurprisingly, the answer is yes.

THEOREM 2.3 *Given any probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a sub- σ -algebra \mathcal{H} of \mathcal{F} and a random variable $X : \Omega \rightarrow \mathbb{R}$, there exist a \mathcal{H} -measurable function $f : \Omega \rightarrow \mathbb{R}$ that satisfies (2.7).*

We close the section by summarizing some additional important properties of conditional expectations. These follow from the definition directly and the reader is invited to prove them in Exercise 2.16, but note the difficulty varies wildly.

THEOREM 2.4 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G} \subset \mathcal{F}$ a sub- σ -algebra of \mathcal{F} , X, Y random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. The following hold true:*

- 1 *If $X \geq 0$, then $\mathbb{E}[X | \mathcal{G}] \geq 0$ almost surely.*
- 2 *$\mathbb{E}[1 | \mathcal{G}] = 1$ almost surely.*
- 3 *$\mathbb{E}[X + Y | \mathcal{G}] = \mathbb{E}[X | \mathcal{G}] + \mathbb{E}[Y | \mathcal{G}]$ almost surely, assuming the expression on the right-hand side is defined.*
- 4 *$\mathbb{E}[XY | \mathcal{G}] = Y\mathbb{E}[X | \mathcal{G}]$ almost surely if $\mathbb{E}[XY]$ exists and Y is \mathcal{G} -measurable.*
- 5 *if $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{F}$, then $\mathbb{E}[X | \mathcal{G}_1] = \mathbb{E}[\mathbb{E}[X | \mathcal{G}_2] | \mathcal{G}_1]$ almost surely.*
- 6 *if \mathcal{G} and \mathcal{F} are independent, then $\mathbb{E}[X | \mathcal{G}] = \mathbb{E}[X]$ almost surely.*
- 7 *If $\mathcal{G} = \{\emptyset, \Omega\}$ is the trivial σ -algebra, then $\mathbb{E}[X | \mathcal{G}] = \mathbb{E}[X]$ almost surely.*

2.7 Notes

- 1 The term σ -algebra (and σ -field) comes from that in many parts of mathematics, the Greek letter σ is the symbol to be used in association with the countably infinites. Countable additivity is in fact often also called σ -additivity. The requirement that additivity should hold for systems of countably infinitely many sets is made so that probabilities (interesting) limiting events are guaranteed to exist.
- 2 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 labeled 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 he/she does not contradict himself/herself). Rational decision makers compare stochastic alternatives based on the alternatives' expected utilities, according to the Von-Neumann-Morgenstern utility theorem. Humans are known to be not doing this, i.e., they are irrational. No surprise here.
- 3 In our toy example instead of $\Omega = [6]^7$, we could have chosen $\Omega = [6]^8$ (considering rolling eight dice instead of 7, one dice never used). There are many other possibilities. We can consider coin flips instead of dice rolls (think about how this could be done). To make this easy, we could use weighted coins (e.g, a coin lands on its head with probability 1/6), but we don't actually need weighted coins (this may be a little tricky to see). The main point is that there are many ways to emulate one randomization device by using another. The difference between these is the set Ω . What makes a choice of Ω viable is if we can emulate the game mechanism on the top of Ω so that in the end the probability of seeing any particular value remains the same. In other words,

the choice of Ω is far from unique. The same is true for the way we calculate the value of the game! For example, the dice could be reordered, if we stay with the first construction. The biggest irony in all probability theory is that we first make a big fuss about introducing Ω and then it turns out that the actual construction of Ω does not matter.

- 4 The Lebesgue σ -algebra is obtained as the completion of the Borel σ -algebra with the following process: Take the **null-sets** in the Borel σ -algebra, which are those the sets with zero Lebesgue measure (one first constructs the Lebesgue measure for the Borel sets). Add all these to the Borel sets and then close the resulting set to make it a σ -algebra. The resulting set is the Lebesgue σ -algebra and the Lebesgue measure is then extended to this set. With the same process, we can complete any σ -algebra with respect to some chosen measure. Incomplete σ -algebras are annoying to work with as one can meet sets that have a zero measure superset but whose measure is not defined.
- 5 How ‘big’ are the Borel and Lebesgue σ -algebras? Can you think of a set that is not Borel measurable? Such sets do exist, but they do not arise naturally in applications. The classic example is called 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 . To reiterate, you do not have to worry much about whether or not functions $X : \mathbb{R} \rightarrow \mathbb{R}$ are Borel. In this book questions of measurability are not related to the fine details of the Borel or Lebesgue σ -algebras. Much more frequently they are related to filtrations and the notion of knowledge available having observed certain random elements.
- 6 We did not talk about this, but there is a whole 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, just to list a few things. The key point is to show first that the composition of measurable maps is a measurable map and that continuous maps are measurable, and then apply these results. For $\limsup_n X_n$, just rewrite it as $\lim_{m \rightarrow \infty} \sup_{n \geq m} X_n$, 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 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.
- 7 The factorization lemma, Lemma 2.1, is attributed to Joseph Doob and Eugene Dynkin. The lemma sneakily uses the properties of real numbers (think about why), so what we said about σ -algebras containing all information is just almost entirely true. The lemma has extensions to more general random elements. In particular, the mathematically oriented readers will find it reassuring that the lemma continues to hold true as long as the σ -algebra of the image space of the random variable to be factored (X in our statement) is a Borel σ -algebra. The key requirement in a way is that this σ -algebra should be rich enough.

- 8 We did not talk about some basic results, like Lebesgue's dominated, monotone convergence theorems, or Fatou's lemma, or Jensen's inequality. Of these, we will definitely use the last, which we elaborate on in the next item. The other results can be found in the texts we cite. These results are concerned with infinite sequence 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 we take expectations before taking limits if we look at asymptotics).
- 9 A simple version of **Jensen's inequality** states that if for any $U \subset \mathbb{R}$ convex, with non-empty interior and for any $f : U \rightarrow \mathbb{R}$ convex function and random variable $X \in U$ such that $\mathbb{E}[X]$ exists, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$. The proof is simple if one notes that for such a convex f function, at every point $m \in R$ in the interior of U , there exists a 'slope' $a \in \mathbb{R}$ such that $a(x-m) + f(m) \leq f(x)$ for all $x \in \mathbb{R}$ (if f is differentiable at m , take $a = f'(m)$). Indeed, if such a slope exists, taking $m = \mathbb{E}[X]$ and replacing x by X we get $a(X - \mathbb{E}[X]) + f(\mathbb{E}[X]) \leq f(X)$. Then, taking the expectation of both sides we arrive at Jensen's inequality. The idea can be generalized into multiple directions, i.e., the domain of f could be a convex set in a vector space, etc.
- 10 The reader might be quite surprised that we have not mentioned densities yet. For most of us our first exposure to probability on continuous spaces was by studying the normal distribution and its density

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2), \quad (2.8)$$

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 $p : \mathbb{R} \rightarrow \mathbb{R}$ is Borel measurable and that the Gaussian measure associated with this density is \mathbb{P} on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by

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

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

$$P(A) = \int_A \frac{dP}{dQ} dQ \quad \text{for all } A \in \mathcal{F}. \quad (2.9)$$

Another word for the Radon-Nikodmm 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 $A \in \Omega$. When $\frac{dP}{dQ}$ exists it follows immediately that P is absolutely continuous with respect to Q by Eq. (2.9). 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 2.5 *Let P, Q be measures on a common measurable space (Ω, \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 (2.9), $f_1 = f_2$ holds Q -almost surely.*

Densities work as expected. Suppose that Z is a standard Gaussian random variable. Without much thinking, we usually write its density as in Eq. (2.8), which we now know is the Radon-Nikodym derivative of the Gaussian measure with respect to the Lebesgue measure. The densities of ‘classical’ distributions are almost always defined with respect to the Lebesgue measure.

- 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 the definition of the densities and the ‘usual machinery’, which means proving the result holds for simple functions and then applying the monotone convergence theorem to take the limit for any measurable function. The chain rule is often used to reduce the calculation of densities to calculation with known densities. 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 ν be the **counting measure** on $(\Omega, 2^\Omega)$, which is defined by $\nu(A) = |A|$. For any P on (Ω, \mathcal{F}) it is easy to see that $P \ll \nu$ and $\frac{dP}{d\nu}(i) = P(\{i\})$, which is sometimes called the distribution function of P .
- 12 The Radon-Nikodym derivative provides one way to define the conditional expectation. Let X be a 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 (Ω, \mathcal{H}) . Define measure μ on (Ω, \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 $\mathbb{E}[X | \mathcal{H}] = \frac{d\mu}{d\mathbb{P}|_{\mathcal{H}}}$ satisfies Eq. (2.7). 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.
- 13 The **Fubini-Tonelli theorem** is a very powerful result that allows one to exchange the order of integrations. This result is needed for example for proving Proposition 2.3 (cf. Exercise 2.15). To state it, we need to introduce **product measurable-spaces** and **product measures**. These work as expected: Given two measurable spaces, $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, the product measurable space is $(\Omega_1 \times \Omega_2, \mathcal{F})$ where \mathcal{F} is the smallest σ -algebra that contains the direct product of \mathcal{F}_1 and \mathcal{F}_2 , defined as $\mathcal{F}_1 \times \mathcal{F}_2 = \{(A_1, A_2) : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}$. We will use $\mathcal{F}_1 \otimes \mathcal{F}_2$ to denote \mathcal{F} and call it the **product σ -algebra**. (The definition can also be extended to the product of multiple measurable spaces. Unsurprisingly the product operation is associative.) Further, given two probability distributions \mathbb{P}_1 and \mathbb{P}_2 on the respective measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, their product measure \mathbb{P} 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 $(A_1, A_2) \in \mathcal{F}_1 \times \mathcal{F}_2$. One can show 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.) Now, the Fubini-Tonelli theorem, which is oftentimes just mentioned as the Fubini theorem, 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 of the product probability space $(\Omega, \mathcal{F}, \mathbb{P}) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \times \mathbb{P}_2)$. Then, if any of the three integrals $\int |X(\omega)| d\mathbb{P}(\omega)$, $\int (\int |X(\omega_1, \omega_2)| d\mathbb{P}_1(\omega_1)) d\mathbb{P}_2(\omega_2)$, $\int (\int |X(\omega_1, \omega_2)| d\mathbb{P}_2(\omega_2)) d\mathbb{P}_1(\omega_1)$ is finite then their values are all equal:

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

- 14 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 functions. 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 is studied in functional analysis. The expectation operator, the way we define it here, is a functional (a special operator) which 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 its linearity, which was stated as a requirement for integrals which defines the \mathbb{E} expectation operator (cf. (2.4)). 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. However, again in line with the previous comment, the abuse is intended and harmless.
- 15 Let $X : \Omega \rightarrow \mathbb{R}$ be a random variable on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The **support** of X is the smallest closed subset $A \subseteq \mathbb{R}$ such that $\mathbb{P}(X \in A) = 1$.

2.8 Bibliographic remarks

Much of this chapter draws inspiration from David Pollard’s “A user’s guide to measure theoretic probability” [Pollard, 2002]. The curious reader should be warned: The notation of the book will be unusual. Pollard follows de Finetti’s notation, where \mathbb{P} and \mathbb{E} gets merged into one. In particular, instead of $\mathbb{E}[X]$, in this notation one writes $\mathbb{P}[X]$. This has the clear advantage that it clearly shows the dependence of the expectation on the probability measure \mathbb{P} (an approximation to this is to use $\mathbb{E}_{\mathbb{P}}[X]$, which is a bit clumsy). Another way of think of this is that all that we have are expectations: In this approach, when writing $\mathbb{P}(A)$ for some event A , we should think of A being replaced by its

indicator so that $\mathbb{P}(A)$ is $\mathbb{E}[\mathbb{I}\{A\}]$. Using indicators in place of sets is also easier at times as we tend to be stronger in algebra than in logic (think of deriving something like $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ with algebra versus with logic). Nevertheless, the main reason we recommend this book is not because of its notation (in fact we realize the notation may be off-putting for some people), but because we find that it is written with extreme care. In particular, this book explains much more about the ‘why’ and ‘how’ than any other book we came across. 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’. We also like the book by [Kallenberg \[2002\]](#). This book is recommended for the mathematically inclined readers who already have a good understanding of the basics: The author has put a major effort into organizing the material so that redundancy is minimized and generality is maximized. This reorganization resulted in quite a few original proofs and the book is comprehensive. The factorization lemma (Lemma 2.1) 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 2.2 is a Corollary of Caratheodory’s extension theorem, which says that probability measures defined on semi-rings have a unique extension to the generated σ -algebra. The remaining results can be found in either of the three books mentioned above. Finally, for something older and less technical we recommend the philosophical essays on probability by Pierre Laplace [[Laplace, 2012](#)].

2.9 Exercises

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

2.2 Let (Ω, \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 $\mathcal{F}_{|A} = \{B \in \mathcal{F} : B \subseteq A\}$.

2.3 Let $\mathcal{G} \subset \Omega$ be an arbitrary nonempty collection of sets and define $\sigma(\mathcal{G})$ as the smallest σ -algebra that contains \mathcal{G} .

- (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 (Ω', \mathcal{F}) is a measurable space and $X : \Omega' \rightarrow \Omega$ 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.

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

2.5 [Bayes law] Verify (2.1).

2.6 Consider the standard probability space $(\Omega, \mathcal{F}, \mathbb{P})$ generated by two standard, unbiased, six-sided dice which 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 any two events, $A \in \sigma(X_1)$ and $B \in \sigma(X_2)$, are independent of each other.

2.7 [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 i th dice roll ($i \in \{1, 2\}$). Show that the events $\{X_1 \leq 2\}$ 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.

2.8 [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) Generalize the result of the previous item to pairwise and mutual independence for collections of events and their indicator random variables.

2.9 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) Let $(\Omega_i, \mathcal{F}_i)$, $i \in \{1, 2, 3\}$ be measurable spaces. Show that if $f_i : \Omega_i \rightarrow \Omega_{i+1}$ is $\mathcal{F}_i/\mathcal{F}_{i+1}$ measurable for $i \in \{1, 2\}$, then their composition, $f_2 \circ f_1 : \Omega_1 \rightarrow \Omega_3$, defined by $(f_2 \circ f_1)(\omega) = f_2(f_1(\omega))$ is $\mathcal{F}_1/\mathcal{F}_3$ -measurable.
- (b) Any $f : \mathbb{R} \rightarrow \mathbb{R}$ continuous function is Borel-measurable.
- (c) Conclude that for any random variable X , $|X|$ is also a random variable.
- (d) Prove that for any random variable X , X is integrable if and only if $|X|$ is integrable. (The statement makes sense since $|X|$ is a random variable whenever X is). Hint: Notice the relationship between $|X|$ and $(X)^+$ and $(X)^-$.

2.10 [Infinite-valued integrals] Can we consistently extend the definition of integrals so that, e.g., for nonnegative random variables, the integral is always defined (it may take on $+\infty$)? Defend your view by either constructing an example (if you are arguing against) or by proving that your definition is consistent with the requirements we have for integrals.

2.11 Prove Proposition 2.1.

2.12 Prove that if $c \in \mathbb{R}$ is a constant, then $\mathbb{E}[cX] = c\mathbb{E}[X]$ (as long as X is integrable).

2.13 Prove Proposition 2.2. Hint: Follow the ‘inductive’ definition of Lebesgue integrals, starting with simple functions, then nonnegative functions and finally arbitrary independent random variables.

2.14 Demonstrate using an example that in general, for dependent random variables, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ does not hold.

2.15 Prove Proposition 2.3. Hint: Consider writing $X(\omega) = \int_{[0,\infty)} \mathbb{I}\{[0, X(\omega)]\}(x) dx$

(why does this hold?) and exchange the integrals. Call the Fubini-Tonelli theorem to justify the exchange of integrals.

2.16 Prove Theorem 2.4.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

3 Stochastic Processes and Markov Chains (†)

The measure-theoretic probability in the previous chapter covers almost all the definitions required. Occasionally, however, we make use of infinite sequences of random variables and for these one requires just a little more machinery. We expect most readers will skip this chapter on the first reading, perhaps referring to it when necessary.

For many applications in this book we only need to construct infinite sequences of independent and identically distributed random variables. In what follows we let λ be the Lebesgue measure on $([0, 1], \mathcal{B}([0, 1]))$. Two measurable spaces $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ are **Borel isomorphic** if there exists an injective function $f : \mathcal{X} \rightarrow \mathcal{Y}$ such that f is \mathcal{X}/\mathcal{Y} -measurable and f^{-1} is \mathcal{Y}/\mathcal{X} -measurable. A **Borel space** is a measurable space $(\mathcal{X}, \mathcal{F})$ that is Borel isomorphic to $(A, \mathcal{B}(A))$ with $A \in \mathcal{B}(\mathbb{R})$ a Borel measurable subset of the reals. The spaces in which random elements usually live are all Borel, including \mathbb{R}^n and its measurable subsets.

THEOREM 3.1 *Let μ be a probability measure on a Borel measurable space \mathcal{S} . Then there exists a sequence of independent random elements X_1, X_2, \dots on $([0, 1], \mathcal{B}([0, 1]), \lambda)$ such that the law $\lambda_{X_t} = \mu$ for all t .*

This allows us to write “let X_1, X_2, \dots be an infinite sequence of independent standard Gaussian random variables” and be comfortable knowing there exists a probability space on which these random variables can be defined. We give a sketch of the proof because, although it is not really relevant for the material in this book, it illustrates the general picture and dispels some of the mystic about what is really going on.

Proof sketch of Theorem 3.1 For simplicity we consider only the case that $\mathcal{S} = ([0, 1], \mathcal{B}([0, 1]))$ and μ is the Lebesgue measure. For any $x \in [0, 1]$ let $F_1(x), F_2(x), \dots$ be the binary expansion of x , which is the unique infinite sequence such that

$$x = \sum_{t=1}^{\infty} F_t(x) 2^{-t}.$$

A direct calculation shows that F_1, F_2, \dots are an infinite sequence of independent Bernoulli random variables. From this we can create an infinite sequence of uniform random variables by reversing the process. To do this we rearrange the

$(F_t)_{t=1}^\infty$ sequence into a grid. For example:

$$\begin{aligned} & F_1, F_2, F_4, F_7, \dots \\ & F_3, F_5, F_8, \dots \\ & F_6, F_9, \dots \\ & F_{10}, \dots \\ & \vdots \end{aligned}$$

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

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 $(\Omega_1 \times \dots \times \Omega_n, \mathcal{F}_1 \otimes \dots \otimes \mathcal{F}_n)$ such that*

$$\mu_{n+1}(A \times \Omega_{n+1}) = \mu_n(A) \quad \text{for all } A \in \Omega_1 \otimes \dots \otimes \Omega_n. \quad (3.1)$$

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



Sequences of measures satisfying Eq. (3.1) are called **projective**.

Theorem 3.2 also demonstrates the existence of an infinite sequence of independent and identically distributed random variables X_1, X_2, \dots with distribution μ . By assumption a random variable takes values in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, 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. Nothing required the measures be identical (or independent), which will sometimes be necessary in the sequel.

3.2 Markov chains

A Markov chain is an infinite sequence of random elements X_1, X_2, \dots where the conditional distribution of X_{t+1} given X_1, \dots, X_t only depends on X_t .

Such random sequences appear throughout probability theory and have many applications besides. The theory is much 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** between $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ is a function $K : \mathcal{X} \times \mathcal{G} \rightarrow [0, 1]$ such that:

- (a) $K(x, \cdot)$ is a measure for all $x \in \mathcal{X}$.
- (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 $Y \sim K(x, \cdot)$. If K_1 is a $(\mathcal{X}, \mathcal{F}) \rightarrow (\mathcal{Y}, \mathcal{G})$ probability kernel and K_2 is a $(\mathcal{Y}, \mathcal{G}) \rightarrow (\mathcal{Z}, \mathcal{H})$ probability kernel, then the **product kernel** $K = K_1 \otimes K_2$ is the probability kernel from $(\mathcal{X}, \mathcal{F}) \rightarrow (\mathcal{Z}, \mathcal{H})$ defined by

$$K(x, A) = \int_{\mathcal{Y}} K_2(y, A) K_1(x, dy),$$

for which an alternate notation is to write $K(x, dz) = \int_{\mathcal{Y}} K_2(y, dz) K_1(x, dy)$. Note the ' d ' appearing inside the measure rather than outside. Occasionally one sees the notation $K_x(A)$ rather than $K(x, A)$, in which case the notation $dK_x(y)$ would make more sense. The product kernel corresponds to taking one step using K_1 followed by a step from K_2 so that $Y \sim K_1(x, \cdot)$ and then $Z \sim K_2(Y, \cdot)$. The counterpart of Theorem 3.2 for Markov chains is known as the Ionescu Tulcea theorem.

THEOREM 3.3 *For each $n \in \mathbb{N}^+$ let $(\Omega_n, \mathcal{F}_n)$ be a measurable space and K_n be a probability kernel from $\prod_{t=1}^{n-1} \Omega_t \rightarrow \Omega_n$. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random elements X_1, X_2, \dots with $X_t : \Omega \rightarrow \Omega_t$ such that $\mathbb{P}_{X_1, \dots, X_n} = \bigotimes_{t=1}^n K_t$ for all $n \in \mathbb{N}^+$.*

A **homogeneous Markov chain** is a sequence of random elements X_1, X_2, \dots taking values in **state space** $\mathcal{S} = (\mathcal{X}, \mathcal{F})$ and with

$$\mathbb{E}[X_{t+1} \in \cdot | X_1, \dots, X_t] = \mathbb{E}[X_{t+1} \in \cdot | X_t] = \mu(X_t, \cdot) \quad \text{almost surely,}$$

where μ is a probability kernel from $(\mathcal{X}, \mathcal{F})$ to $(\mathcal{X}, \mathcal{F})$ and we assume that $\mathbb{E}[X_1 \in \cdot] = \mathbb{P}(X_1 \in \cdot) = \mu_0(\cdot)$ 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 $t \rightarrow \infty$ 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 introduce some of these topics much later in Chapters 36 and 37.

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 $(\mathcal{F}_t)_{t=1}^\infty$ a filtration of \mathcal{F} . We note that for finite sequences the definitions are unchanged except for the obvious range modifications. The sequence $(X_t)_{t=1}^\infty$ is called **\mathcal{F}_t -adapted** if X_t is \mathcal{F}_t -measurable for all t . If (X_t) is \mathcal{F}_t -adapted and for all t the conditional expectation $\mathbb{E}[X_{t+1} | \mathcal{F}_t] = X_t$ almost surely and X_t is integrable, then (X_t) is a **\mathcal{F}_t -martingale**. The dependence on the filtration is often omitted when the underlying filtration is self-evident.

DEFINITION 3.1 A \mathcal{F}_t -adapted sequence of random variables (X_t) is a \mathcal{F}_t -adapted **martingale** if $\mathbb{E}[X_t | \mathcal{F}_{t-1}] = X_{t-1}$ almost surely for all $t \in \{2, 3, \dots\}$. If the equality is replaced with a less/greater-than, then it is called a super/sub-martingale respectively.

EXAMPLE 3.1 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 $\mathcal{F}_t = \sigma(Y_1, \dots, Y_t)$.

DEFINITION 3.2 Let $(\mathcal{F}_t)_{t \in \mathbb{N}}$ be a filtration. A random variable τ with values in $\mathbb{N} \cup \{\infty\}$ is a **stopping time** with respect to (\mathcal{F}_t) if $\mathbb{I}\{\tau \leq t\}$ is \mathcal{F}_t -measurable for all $t \in \mathbb{N}$. The σ -algebra at stopping time τ is

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

A stopping time τ is a random variable that determines when a process stops and that only depends on information available at time τ , which means it cannot ‘peak into the future’ to determine stopping.

EXAMPLE 3.2 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.

THEOREM 3.4 (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 a \mathbb{F} -adapted martingale and τ a \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 $\mathbb{E}[|X_{t+1} - X_t| \mid \mathcal{F}_t] \leq c$ almost surely on the event that $\tau > t$.
- (c) There exists a constant c such that $|X_{t \wedge \tau}| \leq c$ almost surely for all $t \in \mathbb{N}$.

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

One application of Doob's optional stopping theorem is a useful and apriori surprising generalization of Markov's inequality to nonnegative supermartingales.

THEOREM 3.5 (Maximal inequality) *Let $(X_t)_{t=1}^n$ be a supermartingale with $X_t \geq 0$ almost surely for all t . Then*

$$\mathbb{P}\left(\sup_{t \leq n} X_t \geq \varepsilon\right) \leq \frac{1}{\varepsilon} \mathbb{E}[X_1].$$

Proof Let $\tau = (n + 1) \wedge \min\{t \leq n : X_t \geq \varepsilon\}$, where the minimum of an empty set is assumed to be infinite so that $\tau = n + 1$ if $X_t < \varepsilon$ for all $t \in [n]$. Clearly τ is a \mathcal{F}_t -stopping time and $\mathbb{P}(\tau \leq n + 1) = 1$. Then by Theorem 3.4,

$$\mathbb{E}[X_1] \geq \mathbb{E}[X_\tau] \geq \varepsilon \mathbb{P}(\tau \leq n) = \varepsilon \mathbb{P}\left(\sup_{t \leq n} X_t \geq \varepsilon\right),$$

where the second inequality uses the definition of the stopping time and the nonnegativity of the supermartingale. Rearranging completes the proof. \square



Markov's inequality combined with the definition of a supermartingale shows that $\mathbb{P}(X_n \geq \varepsilon) \leq \mathbb{E}[X_1]/\varepsilon$. The maximal inequality is a strict improvement by replacing X_n with $\sup_{t \leq n} X_t$ at no cost whatsoever.

3.4 Notes

- 1 Some authors include in the definition of a stopping time τ that $\mathbb{P}(\tau < \infty) = 1$ and call random times without this property **Markov times**. We do *not* adopt this convention and allow stopping times to be infinite with nonzero probability. Stopping times are also called **optional times**.
- 2 There are several notations for probability kernels depending on the application. The follow 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}_θ is a measure on some measurable space (Ω, \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 $\mathbb{P}(\theta \mid X)$.

- 3 A note on product measure notation.
- 4 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 a little more history on this see the preface of the book by [Meyn and Tweedie \[2012\]](#).

3.5 Bibliographic remarks

There are many places to find the construction of a stochastic process. Like before we recommend [Kallenberg \[2002\]](#) for readers who want to refresh their memory and [Billingsley \[2008\]](#) for a more detailed account. For Markov chains the recent book by [Levin and Peres \[2017\]](#) provides a wonderful introduction. Perhaps followed by the tome of [Meyn and Tweedie \[2012\]](#). A proof of Theorem 3.1 is given Theorem 3.19 in the book by [Kallenberg \[2002\]](#). Theorem 3.2 is credited to Percy John Daniell by [Kallenberg \[2002\]](#) (see [Aldrich 2007](#)). More general versions of this theorem exist. Readers looking for these should look up **Kolmogorov’s extension theorem** [[Kallenberg, 2002, Thm 6.16](#)]. The theorem of Ionescu Tulcea (Theorem 3.3) is attributed to her [[Tulcea, 1949–50](#)] with a modern proof in the book by [[Kallenberg, 2002, Thm 6.17](#)]. There are lots of minor variants of the optional stopping theorem, most of which can be found in any probability book featuring martingales. The most historically notable source is by the man himself [[Doob, 1953](#)]. A more modern book that also gives the maximal inequalities is the book on optimal stopping by [Peskir and Shiryaev \[2006\]](#).

3.6 Exercises

3.1 Fill in the details to prove Theorem 3.1.

3.2 Let X_1, X_2, \dots be an infinite sequence of independent Rademacher random variables and $S_t = \sum_{s=1}^t X_s 2^{s-1}$.

- (a) Show that S_1, S_2, \dots is a martingale.
- (b) Let $\tau = \min\{t : S_t = 1\}$ and show that $\mathbb{P}(\tau < \infty) = 1$.
- (c) What is $\mathbb{E}[S_\tau]$?
- (d) Explain why this does not contradict Doob’s optional stopping theorem.

3.3 Give an example of a martingale S_1, S_2, \dots and stopping time τ such that

$$\lim_{n \rightarrow \infty} \mathbb{E}[X_{\tau \wedge n}] \neq \mathbb{E}[X_\tau].$$

3.4 Show that Theorem 3.5 does not hold in general for supermartingales if the assumption that it be nonnegative is dropped.

3.5 Let τ_1, τ_2, \dots be an almost surely increasing sequence of \mathbb{F} -stopping times on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{F} = (\mathcal{F}_t)$, which means that $\tau_1(\omega) \leq \tau_2(\omega) \leq \dots$ almost surely. Prove that $\tau(\omega) = \lim_{n \rightarrow \infty} \tau_n(\omega)$ is a \mathbb{F} -stopping time.

4 Finite-Armed Stochastic Bandits

A **K -armed stochastic bandit** is a tuple of distributions $\nu = (P_1, P_2, \dots, P_K)$, where P_i is a distribution over the reals for each $i \in [K]$. The learner and the environment interact sequentially as summarized in Fig. 4.1. In each round t the learner chooses action $A_t \in \{1, 2, \dots, K\}$, which is fed to the environment. Then the environment samples reward $X_t \in \mathbb{R}$ from distribution P_{A_t} and reveals it to the learner. The interaction between the learner (or policy) and environment induces a measure on the sequence of outcomes $A_1, X_1, A_2, X_2, \dots, A_n, X_n$ where n is the horizon. Usually the horizon n is finite, but sometimes we allow the interaction to continue indefinitely ($n = \infty$). The interaction diagram above suggests that A_t and X_t should satisfy the following assumptions:

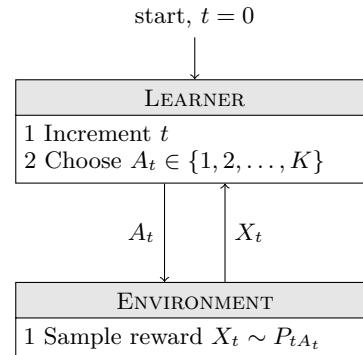


Figure 4.1 Interaction between learner and environment

- (a) The conditional distribution of X_t given $A_1, X_1, \dots, A_{t-1}, X_{t-1}, A_t$ is P_{tA_t} , which captures the intuition that the environment samples X_t from P_{tA_t} in round t .
- (b) $\mathbb{P}(A_t = a | A_1, X_1, \dots, A_{t-1}, X_{t-1}) = \pi_t(A_1, X_1, \dots, A_{t-1}, X_{t-1})$ where π_1, π_2, \dots is a sequence of functions that characterize the learner with $\pi_t(a | a_1, x_1, \dots, a_{t-1}, x_{t-1})$ representing the probability that the learner chooses action a having observed $a_1, x_1, \dots, a_{t-1}, x_{t-1}$. 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 on which probability space A_t and X_t are defined and the measure for which (a) and (b) are satisfied. We show how to do this in Section 4.4, but for now we move on.

4.1 The learning objective

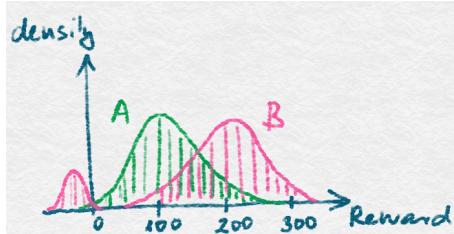
The learner's goal is to maximize the total reward $S_n = \sum_{t=1}^n X_t$, which is a random quantity that depends on the actions of the learner and the rewards sampled by the environment. This is not an optimization problem for three reasons:

- 1 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 .
- 2 The learner does not know the distributions $(P_i)_i$ that determine the reward for each arm.
- 3 What is the value of n for which we are maximizing? 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.

We address the first two points below. For issues relating to knowledge of the horizon it usually suffices to assume the horizon is known when designing a policy. Then if the horizon is unknown the objective is to design a new policy that does not depend on the horizon and is never much worse than what could be achieved for a known horizon. This is almost always quite easy and there exist generic approaches for making the conversion.

Expectation and risk

Suppose that S_n is the revenue of your company. The figure on the right 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 behavior of the cumulative reward, which we will discuss occasionally. In particular, in the situation shown on the above figure, learner B achieves a higher expected total reward than A . However B has a reasonable probability of earning less than A , so a risk sensitive user may prefer learner A .



Environment classes

Even if the horizon is known in advance and we commit to maximizing the expected value of S_n , there is still the problem that $\nu = (P_i)_i$ is unknown. A policy that maximizes the expectation of S_n for one ν can lead to very poor results on another. The learner usually has partial information about the distributions $(P_i)_i$. For example, that the rewards are binary, which means that P_i is Bernoulli

Name	Symbol	Definition
Bernoulli	$\mathcal{E}_{\mathcal{B}}^K$	$\{(\mathcal{B}(\mu_i))_i : \mu \in [0, 1]^K\}$
Uniform	$\mathcal{E}_{\mathcal{U}}^K$	$\{(\mathcal{U}(a_i, b_i))_i : a, b \in \mathbb{R}^K \text{ with } a_i \leq b_i \text{ for all } i\}$
Gaussian (known var.)	$\mathcal{E}_{\mathcal{N}}^K(\sigma^2)$	$\{(\mathcal{N}(\mu_i, \sigma^2))_i : \mu \in \mathbb{R}^K\}$
Gaussian (unknown var.)	$\mathcal{E}_{\mathcal{V}}^K$	$\{(\mathcal{N}(\mu_i, \sigma_i^2))_i : \mu \in \mathbb{R}^K \text{ and } \sigma^2 \in [0, \infty)^K\}$
Finite variance	$\mathcal{E}_{\mathbb{V}}^K(\sigma^2)$	$\{(P_i)_i : \mathbb{V}_{X \sim P_i}[X] \leq \sigma^2 \text{ for all } i\}$
Finite kurtosis	$\mathcal{E}_{\text{Kurt}}^K(\kappa)$	$\{(P_i)_i : \text{Kurt}_{X \sim P_i}[X] \leq \kappa \text{ for all } i\}$
Bounded support	$\mathcal{E}_{[a,b]}^K$	$\{(P_i)_i : \text{Supp}(P_i) \subseteq [a, b]\}$
Subgaussian	$\mathcal{E}_{\text{sg}}^K(\sigma^2)$	$\{(P_i)_i : P_i \text{ is } \sigma\text{-subgaussian for all } i\}$
Supp(P) is the support of distribution P . The kurtosis of a random variable X is a measure of its tail behavior 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 the next chapter.		

Table 4.1 Typical environment classes for stochastic bandits

for each i . We represent this knowledge by defining a set of bandits \mathcal{E} for which $(P_i)_i \in \mathcal{E}$ is guaranteed. Some typical choices 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. For example, by allowing some arms to be Bernoulli and some Gaussian, or have rewards being exponentially distributed, or Gumbel distributed, or belonging to your favorite (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 in fact a natural choice - think of applications like maximizing 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 payoffs 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 payoffs 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 payoff distributions are such that their underlying variance is finite.

Some of the environment classes, like Bernoulli bandits, are **parametric** while others, like subgaussian bandits, are **nonparametric**. The distinction is the number of degrees of freedom needed to describe an element of the environment. 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 $\nu \in \mathcal{E}$. What do we mean by ‘good’? Keep reading! 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 rigorous fashion.

All the environments mentioned so far are **unstructured**, by which we mean that knowledge about the distribution of one arm does not restrict the range of possibilities for other arms. This means the only way to learn about the distribution for an arm is to play it. When we refer to finite-armed stochastic bandits with no further qualifications the reader should take it as assumed that we mean unstructured finite-armed stochastic bandits. Later we will see that much changes in **structured** bandit problems when this property does not hold.

4.2 The regret

In Chapter 1 we informally defined the regret as being the deficit suffered by the learner relative to the optimal policy. Let ν be a K -armed stochastic bandit and define $\mu_i(\nu) = \int_{-\infty}^{\infty} x dP_i(x)$, which is the mean of P_i . Then let $\mu^*(\nu) = \max_{i \in [K]} \mu_i(\nu)$ be the largest mean of all the arms. Of course $\mu_i(\nu)$ could be undefined or infinite, so for the remainder of the book we assume that $\mu_i(\nu)$ exists and is finite for all stochastic bandit instances. For stochastic bandits we define the regret of policy π in bandit ν by

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

where the expectation is taken with respect to the measure on outcomes induced by the interaction of π and ν . Minimizing the regret is equivalent to maximizing the expectation of S_n , but the normalization inherent in the definition of the regret is useful when stating results, which would otherwise need to be stated relative to the optimal action.



If the context is clear we will often drop the dependence on ν and π in various quantities. For example, by writing $R_n = n\mu^* - \mathbb{E}[\sum_{t=1}^n X_t]$. Similarly, when we think readers can work out ranges of symbols in a unique way, we abbreviate sums, or maxima. For example: $\mu^* = \max_i \mu_i$

The regret is always nonnegative and for every bandit ν there exists a policy π for which the regret vanishes.

LEMMA 4.1 *Let ν be a stochastic bandit environment. Then,*

- (a) $R_n(\pi, \nu) \geq 0$ for all policies π .
- (b) The policy π choosing $A_t \in \operatorname{argmax}_i \mu_i$ for all t satisfies $R_n(\pi, \nu) = 0$.
- (c) If $R_n(\pi, \nu) = 0$ for some policy π then for all t , $A_t \in [K]$ is optimal with probability one: $\mathbb{P}(\mu_{A_t} = \mu^*) = 1$.

We leave the proof for the reader (Exercise 4.7). Part (b) of Lemma 4.1 shows that for every bandit ν there exists a policy for which the regret is zero (the best possible outcome). According to Part (c), achieving zero is possible if and only if the learner knows which bandit it is facing (or at least, what is the optimal arm). In general, however, the learner only knows that $\nu \in \mathcal{E}$ for some environment class \mathcal{E} . So what can we hope for? A relatively weak objective is to find a policy π with sublinear regret on all $\nu \in \mathcal{E}$. Formally, this objective is to find a policy π such that

$$\text{for all } \nu \in \mathcal{E}, \quad \lim_{n \rightarrow \infty} \frac{R_n(\pi, \nu)}{n} = 0.$$

If the above holds, then at least the learner is choosing the optimal action almost all of the time as the horizon tends to infinity. One might hope for much more, however. For example that for some specific choice of $C > 0$ and $p < 1$ that

$$\text{for all } \nu \in \mathcal{E}, \quad R_n(\pi, \nu) \leq Cn^p. \quad (4.2)$$

Yet another alternative is to find a function $C : \mathcal{E} \rightarrow [0, \infty)$ and $f : \mathbb{N} \rightarrow [0, \infty)$ such that

$$\text{for all } n \in \mathbb{N}, \nu \in \mathcal{E}, \quad R_n(\pi, \nu) \leq C(\nu)f(n). \quad (4.3)$$

This factorization of the regret into a function of the instance and a function of the horizon is not uncommon in learning theory and appears in particular in supervised learning (for example, Györfi et al. 2002).

We will spend a lot of time in the following chapters finding policies satisfying Eq. (4.2) and Eq. (4.3) for different choices of \mathcal{E} . The form of Eq. (4.3) is quite general, so much time is also spent discovering what are the possibilities for f and C , both of which should be ‘as small as possible’. All of the policies are inspired by the simple observation that in order to make the regret small, the algorithm must discover the action/arm with the largest mean. Usually this means the algorithm should play each arm some number of times to form an estimate of the mean of that arm, and subsequently play the arm with the largest estimated mean. The question essentially boils down to discovering exactly how often the learner must play each arm in order to have reasonable statistical certainty that it has found the optimal arm.

There is another candidate objective called the **Bayesian regret**. If Q is a prior probability measure on \mathcal{E} (which must be equipped with a σ -algebra \mathcal{F}), then the Bayesian regret is the average of the regret with respect to the prior Q .

$$\text{BR}_n(\pi, Q) = \int_{\mathcal{E}} R_n(\pi, \nu) dQ(\nu), \quad (4.4)$$

which is only defined by assuming (or proving) that the regret is a measurable function with respect to \mathcal{F} . An advantage of the Bayesian approach is that having settled on a prior and horizon, the problem of finding a policy that minimizes the Bayesian regret is just an optimization problem. Most of this book is devoted to

analyzing the frequentist regret, but Bayesian methods are covered in Chapters 34 and 35.

4.3 Decomposing the regret

We now present a lemma that forms the basis of almost every proof for stochastic bandits. Let $\nu = (P_i)_{i=1}^K$ be a stochastic bandit and define $\Delta_i(\nu) = \mu^*(\nu) - \mu_i(\nu)$, which is called the **suboptimality gap** or **action gap** or **immediate regret** of action i . Further, let

$$T_i(t) = \sum_{s=1}^t \mathbb{I}\{A_s = i\}$$

be the number of times action i was chosen by the learner after the end of round t . In general, $T_k(n)$ is random, which may seem surprising if we think about a deterministic policy that chooses the same action for any fixed history. So why is $T_k(n)$ random in this case? The reason is because for all rounds t except for the first, the action A_t depends on the rewards observed in rounds $1, 2, \dots, t-1$, which are random, hence A_t will also inherit their randomness. We are now ready to state the second and last lemma of the chapter. In the statement of the lemma we use our convention that the dependence of the various quantities involved on the policy π and the environment ν is suppressed.

LEMMA 4.2 (Regret Decomposition Lemma) *For any policy π and K -armed stochastic bandit environment ν and horizon $n \in \mathbb{N}$, the regret R_n of policy π in ν satisfies*

$$R_n = \sum_{i=1}^K \Delta_i \mathbb{E}[T_i(n)] .$$

The lemma decomposes the regret in terms of the loss due to using each of the arms. It is useful because it tells us that to keep the regret small, the learner should try to minimize the weighted sum of expected action-counts, where the weights are the respective action gaps.



Lemma 4.2 tells us that a learner should aim to use an arm with a larger action gap proportionally fewer times.

Proof of Lemma 4.2 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 $\sum_k \mathbb{I}\{A_t = k\} = 1$. Hence, $S_n = \sum_t X_t = \sum_t \sum_k X_t \mathbb{I}\{A_t = k\}$

and thus

$$R_n = n\mu^* - \mathbb{E}[S_n] = \sum_{k=1}^K \sum_{t=1}^n \mathbb{E}[(\mu^* - X_t)\mathbb{I}\{A_t = k\}] .$$

Now, knowing A_t , the expected reward is μ_{A_t} . Thus we have

$$\begin{aligned} \mathbb{E}[(\mu^* - X_t)\mathbb{I}\{A_t = k\} | A_t] &= \mathbb{I}\{A_t = k\} \mathbb{E}[\mu^* - X_t | A_t] \\ &= \mathbb{I}\{A_t = k\} (\mu^* - \mu_{A_t}) \\ &= \mathbb{I}\{A_t = k\} (\mu^* - \mu_k) . \end{aligned}$$

Using the definition of Δ_k and then plugging in into the right-hand side of the previous equation, followed by using the definition of $T_k(n)$ gives the result. \square

4.4 The canonical bandit model (\dagger)

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 $\mathcal{F}_t = \mathfrak{B}(\mathbb{R}^{2t})|_{\Omega_t}$ be the restriction of the Borel σ -algebra to Ω_t (see Exercise 2.2). The random variables $A_1, X_1, \dots, A_n, X_n$ that make up the outcome are defined by their coordinate projections:

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

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 4.1 A policy π is a sequence π_1, \dots, π_n where π_t is a Markov kernel from $(\Omega_{t-1}, \mathcal{F}_{t-1})$ to $([K], \rho)$ where ρ is the counting measure. Since the latter space is discrete we adopt the notational convention that for $i \in [K]$,

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

Let $\nu = (P_i)_{i=1}^K$ be a stochastic bandit where each P_i is a measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. We want to define a 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 $\pi_t(\cdot | A_1, X_1, \dots, A_{t-1}, X_{t-1})$ 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 4.1.

PROPOSITION 4.1 *Suppose that \mathbb{P} and \mathbb{Q} are measures on an arbitrary measurable space (Ω, \mathcal{F}) and $A_1, X_1, \dots, A_n, X_n$ are random variables on Ω . 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 measure on $(\Omega_n, \mathcal{F}_n)$ that satisfies (a) and (b). To emphasize 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

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

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 rigorous when (P_i) are not discrete requires the use of Radon-Nikodym derivatives. Let λ be a σ -finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ 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 $p_i : \mathbb{R} \rightarrow \mathbb{R}$ such that $\int_B p_i d\lambda = P_i(B)$ for all $B \in \mathcal{B}(\mathbb{R})$. The density $p_{\nu\pi} : \Omega \rightarrow \mathbb{R}$ can now be defined with respect to the product measure $(\rho \times \lambda)^n$ by

$$p_{\nu\pi}(a_1, x_1, \dots, a_n, x_n) = \prod_{t=1}^n \pi(a_t | a_1, x_1, \dots, a_{t-1}, x_{t-1}) p_{a_t}(x_t). \quad (4.5)$$

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

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

It is important to emphasize 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 4.1, 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. For example, the counting measure when (P_i) are discrete and the Lebesgue measure for continuous (P_i) .

There is another way 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 2.2 these random variables may be defined on (Ω, \mathcal{F}) where $\Omega = \mathbb{R}^{nK}$ and $\mathcal{F} = \mathfrak{B}(\mathbb{R}^{nK})$. Then let $X_t = X_{tA_t}$ and A_t is a probability kernel from (Ω, \mathcal{F}_t) to $([K], \rho)$ where $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$. 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. 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, the there does *exists* 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).

4.5 Notes

- 1 Recall that $A \times B$ stands for the Descartes product of sets A and B . Formally, any of the unstructured environments shown in Table 4.1 are of the form $\mathcal{E}^K = (\mathcal{P})^K$ where $(\mathcal{P})^K = \mathcal{P} \times \mathcal{P} \times \dots \times \mathcal{P}$ and \mathcal{P} is some set of distributions over the reals. The upper index in \mathcal{E}^K hints on this and also reminds us that there are K arms. Because of the product form, unstructured environments can also be called ‘product environments’, or ‘rectangle environments’.
- 2 Note that every unstructured environment \mathcal{E}^K is symmetric in the following sense: for any $(P_i)_i \in \mathcal{E}^K$ and any bijection $\pi : [K] \rightarrow [K]$, $(P_{\pi(i)})_i \in \mathcal{E}^K$ also holds. Since a bijection over $[K]$ is also known as a **permutation**, we also say that \mathcal{E}^K is invariant to permutations. While an unstructured environment is necessarily symmetric, a symmetric environment can be structured. Consider for example $K = 2$ and consider the symmetric environment $\mathcal{E} = \{(\mathcal{B}(0), \mathcal{B}(1)), (\mathcal{B}(1), \mathcal{B}(0))\}$. Clearly, \mathcal{E} is symmetric. It is not unstructured, however. If a nonzero reward is observed for the first arm, then the mean of the second arm must be zero.

- 3 While the canonical model introduced in Section 4.4 is enough for finite-armed bandits, in later chapters require similar constructions for more complicated settings. For example, the action space may be infinite or the learner may receive side information that evolves according to a sequence of Markov kernels. In all cases one could construct a canonical model using the same techniques, a task that we leave for connoisseurs of measure theory to tackle for themselves.
- 4 The study of utility and risk has a long history, going right back to (at least) the beginning of probability [Bernoulli, 1954, translated from original Latin, 1738]. The research can broadly be categorized into two branches. The first deals with *describing* how people actually make choices (**descriptive theories**) while the second is devoted to characterizing 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].
- 5 The most famous example of a prescriptive theory is the von Neumann-Morgenstern expected utility theorem, which states that under (reasonable) axioms of rational behavior 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 (that is, rewards). Finance is another field where attitudes toward 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 maximizing 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 minimize 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 2 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 optimization 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.
- 6 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:

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

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

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 penalizing upside risk, but also from failing to capture the skew of the distribution.

- 7 What happens if the distributions of the arms are changing with time? Such bandits are unimaginatively called **nonstationary** bandits. With no assumptions there is not much to be done. Because of this it is usual to assume the distributions change slowly. We’ll eventually see that techniques for stationary bandits can be adapted quite easily to this setup (see Chapter 31).

4.6 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 to the problem 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 was a pathologist who published broadly and apparently did not pursue bandits much further. Sadly his approach was not widely circulated and the algorithm (now called Thompson sampling) did not become popular until very recently. Two decades after Thompson, the bandit problem was formally restated in a short but influential paper by [Robbins \[1952\]](#), an American statistician now most famous for his work on empirical Bayes. Robbin’s introduced the notion of regret and minimax regret in his 1952 paper. The regret decomposition (Lemma 4.2) has been used in practically every work on stochastic bandits and its origin is hard to pinpoint. All we can say for sure is that it does *not* appear in the paper by [Robbins \[1952\]](#), but does appear in the work of [Lai and Robbins \[1985\]](#). [Denardo et al. \[2007\]](#) considers risk in a (complicated) Bayesian setting. [Sani et al. \[2012\]](#) consider a mean-variance approach to risk, while [Maillard \[2013\]](#) considers so-called coherence risk measures (CVaR, is one example of such a risk measure), and with an approach where the regret itself is redefined. Value-at-Risk is considered in the context of a specific bandit policy family by [Audibert et al. \[2007, 2009\]](#).

4.7 Exercises

4.1 Prove Proposition 4.1.

4.2 Prove that the measure defined in terms of the density in Eq. (4.5) satisfies the conditions (a) and (b) in Section 4.4.



Use the properties of the Radon-Nikodym derivative in combination with Fubini's theorem.

4.3 Implement a Bernoulli bandit environment in Python using the code snippet below (or adapt to your favorite language).

```

class BernoulliBandit:
    # accepts a list of K >= 2 floats, each lying in [0,1]
    def __init__(self, means):
        pass

    # Function should return the number of arms
    def K(self):
        pass

    # Accepts a parameter 0 <= a <= K-1 and returns the
    # realisation of random variable X with P(X = 1) being
    # the mean of the (a+1)th arm.
    def pull(self, a):
        pass

    # Returns the regret incurred so far.
    def regret(self):
        pass

```

4.4 Implement the following simple algorithm called ‘Follow-the-Leader’, which chooses each action once and subsequently chooses the action with the largest average observed so far. Ties should be broken randomly.

```

def FollowTheLeader(bandit, n):
    # implement the Follow-the-Leader algorithm by replacing
    # the code below that just plays the first arm in every round
    for t in range(n):
        bandit.pull(0)

```



Depending on the literature you are reading, Follow-the-Leader may be called ‘stay with the winner’ or the ‘greedy algorithm’.

4.5 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

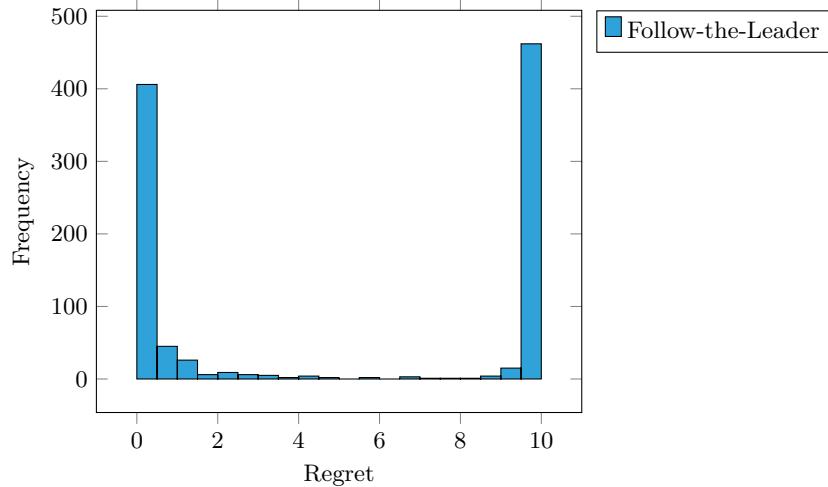


Figure 4.2 Histogram of regret for Follow-the-Leader over 1000 trials on Bernoulli bandit with means $\mu_1 = 0.5, \mu_2 = 0.6$

Follow-the-Leader on the Bernoulli bandit above and record the (random) regret, $n\mu^* - S_n$, in each simulation.

- (b) Plot the results using a histogram. Your figure should resemble Fig. 4.2.
- (c) Explain the results in the figure.

4.6 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 $n \in \{100, 200, 300, \dots, 1000\}$.
- (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?

4.7 Prove Lemma 4.1.



All items follow from Lemma 4.2.

4.8 Suppose ν is a finite-armed stochastic bandit and π is a policy such that

$$\lim_{n \rightarrow \infty} \frac{R_n(\pi, \nu)}{n} = 0.$$

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

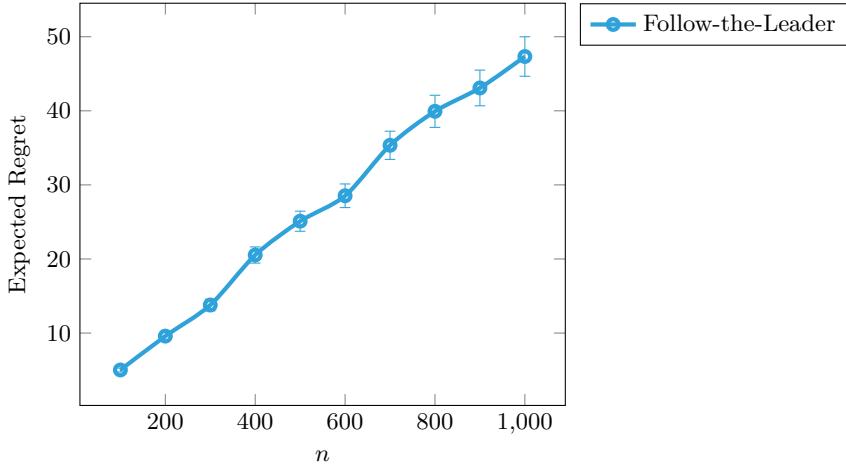


Figure 4.3 Histogram of regret for Follow-the-Leader over 1000 trials on a Bernoulli bandit with means $\mu_1 = 0.5, \mu_2 = 0.6$

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

4.9 [One-armed bandits] This exercise is concerned with a very simple model called the **one-armed bandit**. A bar contains a single slot machine. Playing costs \$1 and the payoff is either \$2 or \$0 with probabilities p and $1-p$ respectively. Of course you do not know p and – unlike in the real world – we will assume it could reasonably take on any value in $[0, 1]$. The game proceeds over n rounds, where in each round you choose either to play the machine or do nothing. If you do nothing, then your reward is $X_t = 0$. If you play the machine, then your reward is $X_t = 1$ with probability p and $X_t = -1$ otherwise. A policy in this case chooses either PLAY or DO NOTHING based on the history. The expected regret of policy π is given by

$$R_n(p, \pi) = n \max \{0, 2p - 1\} - \mathbb{E} \left[\sum_{t=1}^n X_t \right],$$

where X_1, \dots, X_n are the random rewards earned by π .

- (a) Describe an optimal policy when p is known (your policy should depend on p).
- (b) A policy is called a **retirement policy** if it chooses to play the machine until some (possibly random) time and then does nothing until the game ends. Prove that if n is known, then for any policy π there exists a retirement policy π' such that

$$R_n(p, \pi') \leq R_p^\pi(n) \text{ for all } p.$$

- (c) Prove that if n is not known, then all retirement policies have linear regret for some $p \in [0, 1]$ as n tends to infinity.



For (b) specify what the policy π' does given that it has access to the policy π . One easy way of doing this is assuming that π has a memory of past observations it has two subroutines; one for getting the next action and one for feeding π with the next observation. Show in a pseudocode how π' would use π through these two subroutines and argue that π' is indeed a retirement policy. For (c) use that in a stochastic bandit problem the regret can be written as $R_n = \sum_{i: \Delta_i > 0} \Delta_i \mathbb{E}[T_i(n)]$ where Δ_i are the action gaps and $T_i(n)$ is the number of times arm i is chosen.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

5 Concentration of Measure

Before we can start designing and analyzing 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 payoffs are initially unknown, they must be learned from data. We now ask how long it takes to learn about the mean reward of an action.

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 define an **estimator** of the common mean μ . The natural choice to estimate μ is to use the average of the observations, also known as the **sample mean** or **empirical mean**.

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

The question is how far from μ do we expect $\hat{\mu}$ to be? First, by the linearity of expectation (Proposition 2.1), we notice that $\mathbb{E}[\hat{\mu}] = \mu$. A simple measure of the spread of the distribution of a random variable Z is its variance, $\mathbb{V}[Z] = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$. A quick calculation using independence shows that $\mathbb{V}[\hat{\mu}] = \sigma^2/n$. From this we get

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

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 (so larger variance means larger expected squared difference). 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 analyze 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 ?

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

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

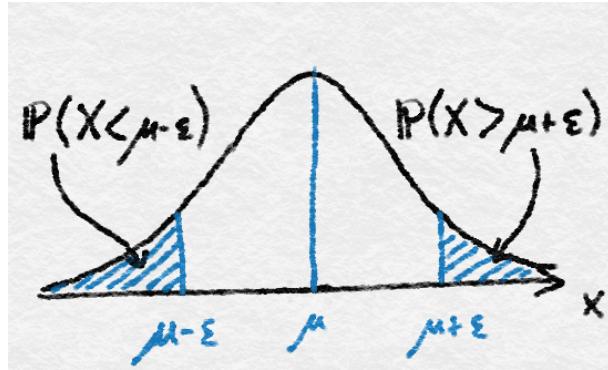


Figure 5.1 The figure shows a probability density, with the tails shaded indicating the regions where X is more than ε away from the mean μ .

5.1 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 with finite mean and $\varepsilon > 0$ it holds that:*

- (a) (*Markov*): $\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}[|X|]}{\varepsilon}$.
- (b) (*Chebyshev*): *If $\mathbb{V}[X] < \infty$, then $\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathbb{V}[X]}{\varepsilon^2}$.*

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

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

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 in many cases the inequality is extremely loose and that huge improvement is possible if the distribution of X is well behaved. In particular, by assuming that higher moments of X exist, Chebyshev's inequality can be greatly 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 optimized. This is a bit cumbersome and thus instead we present the continuous analog of this, known as the Cramer-Chernoff method.

To calibrate our expectations on what gains to expect over Chebyshev's inequality, let us first discuss the **central limit theorem**. Let $S_n = \sum_{t=1}^n (X_t - \mu)$.

The central limit theorem (CLT) says that under no additional assumptions than the existence of the variance, the limiting distribution of $S_n/\sqrt{\sigma^2 n}$ as $n \rightarrow \infty$ is a Gaussian with mean zero and unit variance. Now, if $Z \sim \mathcal{N}(0, 1)$,

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

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

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

which gives

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

This is usually much smaller than what we obtained with Chebyshev's inequality (cf. 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. Unfortunately, since the central limit theorem is asymptotic, we cannot use it to study the regret when the number of rounds is a fixed finite number (cf. Exercise 5.5). Despite the folk-rule that $n = 30$ is sufficient for the Gaussian approximation based on the CLT to be reasonable, this is simply not true. One well known example is provided by Bernoulli variables with parameter $p \approx 1/n$, in which case the distribution of the sum is known to be much better approximated by the Poisson distribution with parameter one, which is nowhere near similar to the Gaussian distribution.

For these reasons we need a non-asymptotic alternative to the CLT. The pathological example in the previous paragraph shows this is only possible by making some additional assumptions.

5.2 The Cramer-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 5.1 (Subgaussianity) A random variable X is σ -subgaussian if for all $\lambda \in \mathbb{R}$ it holds that $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2 \sigma^2 / 2)$.

An alternative way to express the subgaussianity condition uses the **moment**

generating function of X , which is a function $M_X : \mathbb{R} \rightarrow \mathbb{R}$ defined by $M_X(\lambda) = \mathbb{E}[\exp(\lambda X)]$. The condition in the definition can be written as

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

Another useful function is the **cumulative generating function**, which is a function $\psi_X : \mathbb{R} \rightarrow \mathbb{R}$ defined by $\psi_X(\lambda) = \log M_X(\lambda)$. The origin of the names of M_X and ψ_X are explained in the notes and Exercise 5.9. 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

$$\mathbb{E}[\exp(\lambda X)] = \int_0^\infty \underbrace{\exp(-x)}_{\text{density of exponential}} \times \exp(\lambda x) dx = \infty.$$

Therefore the definition of a subgaussian places a nontrivial restriction on the random variables by assuming that the domain of the moment generating function is the whole real line. It is not hard to verify that the moment generating function of a zero-mean Gaussian with variance σ^2 is $\exp(\lambda^2 \sigma^2 / 2)$ from which we conclude that a centered Gaussian with standard deviation $\sigma > 0$ is σ -subgaussian.

Where does the term ‘subgaussian’ come from? The following result provides the explanation. The tails of a σ -subgaussian random variable decay approximately as fast as that of a Gaussian with zero mean and the same variance

THEOREM 5.1 *If X is σ -subgaussian, then for any $\varepsilon \geq 0$,*

$$\mathbb{P}(X \geq \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right). \quad (5.5)$$

Proof We take a generic approach called **Cramer-Chernoff’s method**. Let $\lambda > 0$ be some constant to be tuned later. Then

$$\begin{aligned} \mathbb{P}(X \geq \varepsilon) &= \mathbb{P}(\exp(\lambda X) \geq \exp(\lambda \varepsilon)) \\ &\leq \mathbb{E}[\exp(\lambda X)] \exp(-\lambda \varepsilon) \quad (\text{Markov's inequality}) \\ &\leq \exp\left(\frac{\lambda^2 \sigma^2}{2} - \lambda \varepsilon\right). \quad (\text{Def. of subgaussianity}) \end{aligned}$$

Now λ was any positive constant, and in particular may be chosen to minimize the bound above, which is achieved by $\lambda = \varepsilon/\sigma^2$. \square

A similar inequality holds for the left tail. By using the union bound $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ we also find that $\mathbb{P}(|X| \geq \varepsilon) \leq 2 \exp(-\varepsilon^2/(2\sigma^2))$. An equivalent form of these bounds is:

$$\mathbb{P}\left(X \geq \sqrt{2\sigma^2 \log(1/\delta)}\right) \leq \delta \quad \mathbb{P}\left(|X| \geq \sqrt{2\sigma^2 \log(2/\delta)}\right) \leq \delta.$$

This form is often more convenient and especially the latter, which for small δ shows that with overwhelming probability the random variable X takes values in

the interval

$$\left(-\sqrt{2\sigma^2 \log(2/\delta)}, \sqrt{2\sigma^2 \log(2/\delta)} \right).$$

To study the tail behavior of $\hat{\mu} - \mu$, we need one more lemma (the first item of the lemma is included for completeness):

LEMMA 5.2 *Suppose that X is σ -subgaussian and X_1 and X_2 are independent and σ_1 and σ_2 -subgaussian respectively, then:*

- (a) $\mathbb{E}[X] = 0$ and $\mathbb{V}[X] \leq \sigma^2$.
- (b) cX is $|c|\sigma$ -subgaussian for all $c \in \mathbb{R}$.
- (c) $X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$ -subgaussian.

The proof of the lemma is left to the reader (Exercise 5.7). Note that if X_1 and X_2 are *not* independent then $X_1 + X_2$ is still guaranteed to be $(\sigma_1 + \sigma_2)$ -subgaussian. The difference between this and $\sqrt{\sigma_1^2 + \sigma_2^2}$ is the price of losing independence.

With this we are ready for our key concentration inequality. In particular, combining Lemma 5.2 and Theorem 5.1 leads to a very straightforward analysis of the tails of $\hat{\mu} - \mu$ under the assumption that $X_i - \mu$ are σ -subgaussian. Since X_i are assumed to be independent, by the lemma it holds that $\hat{\mu} - \mu = \sum_{i=1}^n (X_i - \mu)/n$ is σ/\sqrt{n} -subgaussian.

COROLLARY 5.1 *Assume that $X_i - \mu$ are independent, σ -subgaussian random variables. Then for any $\varepsilon \geq 0$*

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) \quad \text{and} \quad \mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right),$$

where $\hat{\mu} = \frac{1}{n} \sum_{t=1}^n X_t$.

By the inequality $\exp(-x) \leq 1/(ex)$ (which holds for all $x \geq 0$) we can see that except for a very small ε the above inequality is strictly stronger than what we obtained via Chebyshev's inequality and exponentially smaller (tighter) if $n\varepsilon^2$ is large relative to σ^2 .

The alternative deviation form of the above result says that under the conditions of the result, for any $\delta \in [0, 1]$, with probability at least $1 - \delta$,

$$\mu \leq \hat{\mu} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}. \tag{5.6}$$

Symmetrically, it also follows that with probability at least $1 - \delta$,

$$\mu \geq \hat{\mu} - \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}. \tag{5.7}$$

Again, one can use a union bound to derive a two-sided inequality.

Before we finally return to bandits, one might be wondering what variables are subgaussian? We give three basic examples. First, as was already mentioned, if X is distributed like a Gaussian with zero mean and variance σ^2 , then X is

σ -subgaussian. Second, if X is bounded, zero-mean (i.e., $\mathbb{E}[X] = 0$ and $|X| \leq B$ almost surely for some $B \geq 0$) then X is B -subgaussian. A special case is when X is a shifted Bernoulli with $\mathbb{P}(X = 1 - p) = p$ and $\mathbb{P}(X = -p) = 1 - p$. In this case it also holds that X is $1/2$ -subgaussian. Finally, recall that the exponential serves the role of a classical distribution that is *not* subgaussian (instead it is sub-exponential, but we will not concern ourselves with this).



For random variables that are not **centered** ($\mathbb{E}[X] \neq 0$) we will abuse notation by saying that X is σ -subgaussian if the **noise** $X - \mathbb{E}[X]$ is σ -subgaussian. A distribution is called σ -subgaussian if a random variable drawn from that distribution is σ -subgaussian. In fact, the subgaussianity property is really a property of both a random variable and the measure on the space on which it is defined, so the nomenclature is doubly abused.

5.3 Notes

- 1 The Berry-Esseen Theorem (independently discovered by [Berry \[1941\]](#) and [Esseen \[1942\]](#)) quantifies the speed of convergence in the CLT. It essentially says that the distance between the Gaussian and the actual distribution decays at a rate of $1/\sqrt{n}$ under some mild assumptions (see Exercise 5.5). This is known to be tight for the class of probability distributions that appear in the Berry-Esseen result. However, this is a vacuous result when the tail probabilities themselves are much smaller than $1/\sqrt{n}$. Hence the need for concrete finite-time results.
- 2 Theorem 5.1 shows that subgaussian random variables have tails that decay almost as fast as a Gaussian. A version of the converse is also possible. That is, if a centered random has tails that behave in a similar way to a Gaussian, then it is subgaussian. In particular, the following holds: Let X be a centered random variable ($\mathbb{E}[X] = 0$) with $\mathbb{P}(|X| \geq \varepsilon) \leq 2 \exp(-\varepsilon^2/2)$. Then X is

$\sqrt{5}$ -subgaussian:

$$\begin{aligned}
 \mathbb{E}[\exp(\lambda X)] &= \mathbb{E}\left[\sum_{i=0}^{\infty} \frac{\lambda^i X^i}{i!}\right] \leq 1 + \sum_{i=2}^{\infty} \mathbb{E}\left[\frac{\lambda^i |X|^i}{i!}\right] \\
 &\leq 1 + \sum_{i=2}^{\infty} \int_0^{\infty} \mathbb{P}\left(|X| \geq \frac{i!^{1/i}}{\lambda} x^{1/i}\right) dx \quad (\text{Exercise 2.15}) \\
 &\leq 1 + 2 \sum_{i=2}^{\infty} \int_0^{\infty} \exp\left(-\frac{i!^{2/i} x^{2/i}}{2\lambda^2}\right) dx \quad (\text{by assumption}) \\
 &= 1 + \sqrt{2\pi}\lambda \left(\exp(\lambda^2/2) \left(1 + \operatorname{erf}\left(\frac{\lambda}{\sqrt{2}}\right)\right) - 1 \right) \quad (\text{by Mathematica}) \\
 &\leq \exp\left(\frac{5\lambda^2}{2}\right).
 \end{aligned}$$

This bound is surely loose. At the same time, there is little room for improvement: If X has density $p(x) = |x| \exp(-x^2/2)/2$, then $\mathbb{P}(|X| \geq \varepsilon) = \exp(-\varepsilon^2/2)$. And yet X is at best $\sqrt{2}$ -subgaussian, so some degree of slack is required (see Exercise 5.4).

- 3 The classical CLT only applies to sequences of independent and identically distributed random variables with finite variance. It turns out that these conditions can be relaxed significantly. One such relaxation is the removal of the condition that the sequence be identically distributed. A CLT-like result still holds under **Lindeberg's condition**, which ensures that the variance is not caused by increasingly infrequent (and catastrophic events). Formally, let $(X_t)_t$ be a sequence of independent random variables with means $(\mu_t)_t$ and variances $(\sigma_t^2)_t$ and let $s_n^2 = \sum_{t=1}^n \sigma_t^2$. Then the Lindeberg CLT says that if for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \frac{1}{s_n^2} \sum_{t=1}^n \mathbb{E}[(X_t - \mu_t)^2 \mathbb{I}\{|X_t - \mu_t| \geq \varepsilon s_n\}] = 0,$$

then the random variable given by $Z_n = \frac{1}{s_n} \sum_{t=1}^n (X_t - \mu_t)$ converges in distribution to a standard normal distribution. We have little use for this theorem as-is because like the CLT, it only holds asymptotically. It does, however, provide inspiration for what might be possible in finite-time and we will see similar results in subsequent chapters. The interested reader can find much more on this in the classic text by [Billingsley \[2008\]](#), which includes many other generalizations such as the multi-variate case.

- 4 We saw in (5.4) that if X_1, X_2, \dots, X_n are independent standard Gaussian random variables and $\hat{\mu} = \frac{1}{n} \sum_{t=1}^n$, then

$$\mathbb{P}(\hat{\mu} \geq \varepsilon) \leq \sqrt{\frac{\sigma^2}{2\pi n\varepsilon^2}} \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

If $n\varepsilon^2/\sigma^2$ is relatively large, then this bound is marginally stronger than $\exp(-n\varepsilon^2/(2\sigma^2))$ that follows from the subgaussian analysis. One might ask

whether or not a similar improvement is possible more generally. And [Talagrand \[1995\]](#) will tell you: Yes! At least for bounded random variables (details in the paper).

- 5 The name ‘moment generating function’ comes from the following fact. Suppose that $M_X(\lambda)$ exists in a neighborhood of zero, then all the moments of the underlying random variable can be read out from its derivatives at zero (see Exercise 5.9). Furthermore, the moment generating function in any small neighborhood of zero uniquely determines the distribution of the underlying random variable. In particular, the following holds: Let X and Y be random variables and $a > 0$ and assume that $\text{dom}(M_X) \supseteq [-a, a]$ and $\text{dom}(M_Y) \supseteq [-a, a]$ and $M_X(\lambda) = M_Y(\lambda)$ on $[-a, a]$. Then
- (e) X and Y have the same distribution: $\mathbb{P}(X \geq x) = \mathbb{P}(Y \geq x)$ for all $x \in \mathbb{R}$.
 - (e) Suppose additionally that X_1, X_2, \dots are a sequence of random variables such that $\text{dom}(M_{X_t}) \supseteq [-a, a]$ and $\lim_{t \rightarrow \infty} M_{X_t}(\lambda) = M_X(\lambda)$ for all $\lambda \in [-a, a]$. Then $\lim_{t \rightarrow \infty} \mathbb{P}(X_t \geq x) = \mathbb{P}(X \geq x)$ for all x , which is equivalent to saying that $(X_t)_t$ converges in distribution to X .

The proof of this result does not belong here (but could serve as a challenging exercise). Most probability texts prove the analogous result for the characteristic function (see the next note), which is known as **Lévy’s continuity theorem**, but the above is sometimes also given [[Billingsley, 2008](#), §30]. The significance of these results is that the moment generating function is often more convenient to work with than the distribution. For example if X and Y are independent, then the distribution of $X + Y$ is the convolution of the distributions of X and Y , which, in a way, is a complicated object. The moment generating function, on the other hand, satisfies $M_{X+Y}(\lambda) = M_X(\lambda)M_Y(\lambda)$. To illustrate the usefulness of this, let X, X_1, X_2, \dots, X_n be a sequence of independent random variables with zero mean, unit variance and $M_X(\lambda)$ defined for all $\lambda \in [-a, a] \subset \mathbb{R}$ with $a > 0$ and let $Z_n = \sum_{t=1}^n X_t / \sqrt{n}$. By the multiplicative property above, for any $\lambda \in [-a, a]$, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \log M_{Z_n}(\lambda) &= \lim_{n \rightarrow \infty} n \log M_X(\lambda / \sqrt{n}) \\ &= \lim_{n \rightarrow \infty} n \log \mathbb{E}[\exp(\lambda X / \sqrt{n})] \\ &= \lim_{n \rightarrow \infty} n \log \left(1 + \frac{\lambda^2}{2n} + \sum_{i=3}^{\infty} \frac{\lambda^i \mathbb{E}[X^i]}{i! n^{i/2}} \right) \\ &= \frac{\lambda^2}{2} \\ &= \log M_Z(\lambda), \end{aligned}$$

where Z is distributed like a standard Gaussian. Therefore the sequence Z_n converges in distribution to Z , which is exactly the statement of the central limit theorem. Of course here we required the additional assumption that $M_X(\lambda)$ was defined over $[-a, a]$ with $a > 0$, which does not normally appear in the statement of the CLT.

- 6 The non-existence of the moment generating function is one of the motivations to introduce the **characteristic function**, which is defined as $\phi_X(\lambda) = \mathbb{E}[\exp(\lambda iX)]$ with $i = \sqrt{-1}$ being the imaginary unit and which always exists and shares many properties with the moment generating function. An example application of characteristic functions is the classical proof of the central limit theorem. Mathematically inclined readers will notice that the moment generating function of random variable X is the Laplace transform of $-X$ (and the characteristic function is the Fourier transform). There are many books on these topics, so we'll just mention that they are essential tools for a probabilist and leave it at that.
- 7 Hoeffding's lemma states that for a zero-mean random variable X such that $X \in [a, b]$ almost surely for real values $a < b$, then $M_X(\lambda) \leq \exp(\lambda^2(b-a)^2/8)$. Applying Chernoff's method shows that if X_1, X_2, \dots, X_n are independent and $X_t \in [a_t, b_t]$ almost surely with $a_t < b_t$ for all t , then

$$\mathbb{P}\left(\sum_{t=1}^n (X_t - \mathbb{E}[X_t]) \geq \varepsilon\right) \leq \exp\left(\frac{2n^2\varepsilon^2}{\sum_{t=1}^n (b_t - a_t)^2}\right). \quad (5.8)$$

For details see Exercise 5.13. There are many variants of this result that provide tighter bounds when X satisfies certain additional distributional properties. For example, if X has small variance, then **Bernstein's inequality** supplies a useful improvement. For details see the texts mentioned below.

- 8 The Cramer-Chernoff method is applicable beyond the subgaussian case, even when the moment generating function is not defined globally. One example where this occurs is when X_1, X_2, \dots, X_n are independent standard Gaussian and $Y = \sum_{i=1}^n X_i^2$. Then Y has a χ^2 -distribution with n degrees of freedom. An easy calculation shows that $M_Y(\lambda) = (1 - 2\lambda)^{-n/2}$ for $\lambda \in [0, 1/2)$ and $M_Y(\lambda)$ is undefined for $\lambda \geq 1/2$. By the Cramer-Chernoff method we have

$$\begin{aligned} \mathbb{P}(Y \geq n + \varepsilon) &\leq \inf_{\lambda \in [0, 1/2)} M_\lambda(Y) \exp(-\lambda(n + \varepsilon)) \\ &\leq \inf_{\lambda \in [0, 1/2)} \left(\frac{1}{1 - 2\lambda}\right)^{\frac{n}{2}} \exp(-\lambda(n + \varepsilon)) \end{aligned}$$

Choosing $\lambda = \frac{1}{2} - \frac{n}{2(n+\varepsilon)}$ leads to $\mathbb{P}(Y \geq n + \varepsilon) \leq \left(1 + \frac{\varepsilon}{n}\right)^{\frac{n}{2}} \exp\left(-\frac{\varepsilon}{2}\right)$, which turns out to be about the best you can do [Laurent and Massart, 2000].

- 9 Distributions for which the moment generating function is infinite for all $\lambda > 0$ are called **heavy tailed**. Distributions that are not heavy tailed are **light tailed**.

5.4 Bibliographical remarks

We will be returning to concentration of measure many times, but note here that it is an interesting (and still active) topic of research. What we have seen is only the tip of the iceberg. Readers that are interested to dive into this exciting

field might enjoy the book by [Boucheron et al. \[2013\]](#). For matrix versions of many standard results there is a recent book by [Tropp \[2015\]](#). The survey of [McDiarmid \[1998\]](#) has many of the classic results. We'll often see concentration of the empirical mean for random variables that are not quite independent and very far from identically distributed. There is a useful type of concentration bound that are ‘self-normalized’ by the variance. A nice book on this is by [Peña et al. \[2008\]](#). Another tool that is occasionally useful for deriving concentration bounds in more unusual setups is called **empirical process theory**. There are several references for this, including those by [van de Geer \[2000\]](#) or [Dudley \[2014\]](#). Of course these are just a few of the many textbooks (not to mention the papers). We will return to concentration many times throughout the book, so more details will follow, especially on martingales.

5.5 Exercises

There are too many candidate exercises to list. We heartily recommend *all* the exercises in Chapter 2 of the book by [Boucheron et al. \[2013\]](#).

5.1 Let X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables with mean μ and variance $\sigma^2 < \infty$. Let $\hat{\mu} = \frac{1}{n} \sum_{t=1}^n X_t$ and show that $\mathbb{V}[\hat{\mu}] = \mathbb{E}[(\hat{\mu} - \mu)^2] = \sigma^2/n$.

5.2 Prove Markov’s inequality (Lemma 5.1).

5.3 Compare the Gaussian tail probability bound on the right-hand side of (5.4) and the one on (5.2). What values of ε make one smaller than the other? Discuss your findings.

5.4 Let X be a random variable on \mathbb{R} with density with respect to the Lebesgue measure of $p(x) = |x| \exp(-x^2/2)/2$. Show that:

- (a) $\mathbb{P}(|X| \geq \varepsilon) = \exp(-\varepsilon^2/2)$.
- (b) That X is not $\sqrt{(2-\varepsilon)}$ -subgaussian for any $\varepsilon > 0$.

5.5 Let X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables with mean μ , variance σ^2 and bounded third absolute moment

$$\rho = \mathbb{E}[|X_1 - \mu|^3] < \infty.$$

Let $S_n = \sum_{t=1}^n (X_t - \mu)/\sigma$. The Berry-Esseen theorem shows that

$$\left| \mathbb{P}\left(\frac{S_n}{\sqrt{n}} \geq x\right) - \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-y^2/2) dy}_{\Phi(x)} \right| \leq \frac{C\rho}{\sqrt{n}},$$

where $C < 1/2$ is a universal constant.

- (a) Let $\hat{\mu}_n = \frac{1}{n} \sum_{t=1}^n X_t$ and derive a tail bound from the Berry-Esseen theorem. That is, give a bound of the form $\mathbb{P}(\hat{\mu}_n \geq \mu + \varepsilon)$ for positive values of ε .
- (b) Compare your bound with the one that can be obtained from the Cramer-Chernoff method. Argue pro- and contra- for the superiority of one over the other.

5.6 We mentioned that invoking the central limit theorem to approximate the distribution of sums of independent Bernoulli random variables using a Gaussian can be a bad idea. Let $X_1, \dots, X_n \sim \mathcal{B}(p)$ be independent Bernoulli random variables with common mean $p = \lambda/n$ where $\lambda \in (0, 1)$. For $x \in \mathbb{N}$ natural number, let $P_n(x) = \mathbb{P}(X_1 + \dots + X_n = x)$.

- (a) Show that $\lim_{n \rightarrow \infty} P_n(x) = e^{-\lambda} \lambda^x / (x!)$, which is a Poisson distribution with parameter λ .
- (b) Explain why this does not contradict the CLT and discuss the implications of the Berry-Esseen.
- (c) In what way does this show that the CLT is indeed a poor approximation in some cases?
- (d) Based on Monte-Carlo simulations, plot the distribution of $X_1 + \dots + X_n$ for $n = 30$ and some well-chosen values of λ . Compare the distribution to what you would get from the CLT. What can you conclude?

5.7 Prove Lemma 5.2.



Use Taylor series.

5.8 Let X_i be σ_i -subgaussian for $i \in \{1, 2\}$ with $\sigma_i \geq 0$. Prove that $X_1 + X_2$ is $(\sigma_1 + \sigma_2)$ -subgaussian. Do *not* assume independence of X_1 and X_2 .

5.9 [Properties of moment/cumulant generating functions] Let X be a real-valued random variable and let $M_X(\lambda) = \mathbb{E}[\exp(\lambda X)]$ be its moment-generating function defined over $\text{dom}(M_X) \subset \mathbb{R}$ where the expectation takes on finite values. Show that the following properties hold:

- (a) M_X is convex and in particular $\text{dom}(M_X)$ is an interval containing zero.
- (b) $M_X(\lambda) \geq e^{\lambda \mathbb{E}[X]}$ for all $\lambda \in \text{dom}(M_X)$.
- (c) For any λ in the interior of $\text{dom}(M_X)$, M_X is infinitely many times differentiable.
- (d) Let $M_X^{(k)}(\lambda) = \frac{d^k}{d\lambda^k} M_X(\lambda)$. Then, for λ in the interior of $\text{dom}(M_X)$, $M^{(k)}(\lambda) = \mathbb{E}[X^k \exp(\lambda X)]$.
- (e) Assuming 0 is in the interior of $\text{dom}(M_X)$, $M_X^{(k)}(0) = \mathbb{E}[X^k]$ (hence the name of M_X).
- (f) ψ_X is convex (that is, M_X is log-convex).



For part (a) use the convexity of $x \mapsto e^x$.

5.10 Let X, X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables with zero mean and moment generating function M_X with $\text{dom}(M_X) = \mathbb{R}$. Let $\hat{\mu}_n = \frac{1}{n} \sum_{t=1}^n X_t$.

(a) Show that for any $\varepsilon > 0$,

$$\frac{1}{n} \log \mathbb{P}(\hat{\mu}_n \geq \varepsilon) \leq -\psi_X^*(\varepsilon) = -\sup_{\lambda} (\lambda \varepsilon - \log M_X(\lambda)). \quad (5.9)$$

(b) Let $\sigma^2 = \mathbb{V}[X]$. The central limit theorem says that for any $x \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\hat{\mu}_n \sqrt{\frac{n}{\sigma^2}} \geq x\right) = \Phi(x),$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-y^2/2) dy$ is the cumulative distribution of the standard Gaussian. Let Z be a random variable distributed like a standard Gaussian. A careless application of this result might suggest that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\hat{\mu}_n \geq \varepsilon) \stackrel{?}{=} \lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}\left(Z \geq \varepsilon \sqrt{\frac{n}{\sigma^2}}\right).$$

Evaluate the right-hand side and explain why the question-marked equality does *not* hold.



As it happens, the inequality in (5.9) may be replaced by an equality as $n \rightarrow \infty$. The assumption that the moment-generating function exists everywhere may be relaxed significantly. We refer the interested reader to the classic text by [Dembo and Zeitouni \[2009\]](#). The function ψ_X^* is called the **Legendre transform**, **convex conjugate** or **Fenchel dual** of the convex function ψ_X . Convexity will play a role in some of the later chapters and will be discussed in more detail then. The standard reference is by [Rockafellar \[2015\]](#).

5.11 A **Rademacher** random variable X satisfies $X \in \{-1, 1\}$ with $\mathbb{P}(X = 1) = 1/2$ and $\mathbb{P}(X = -1) = 1/2$. Prove that $M_X(\lambda) = \cosh(\lambda) \leq \exp(\lambda^2/2)$.

5.12 Prove that if X is zero-mean and $a \leq X \leq b$ almost surely for some $a < 0 < b$, then X is $(b - a)$ -subgaussian.



Let X' be an independent copy of X and consider $M_{X-X'}$. Use $M_X \geq 1$, which follows from Exercise 5.9 and then use $M_{X-X'} = M_{\varepsilon(X-X')}$ where $\varepsilon \in \{-1, +1\}$ is a Rademacher random variable independent of (X, X') . Then use the result of the previous exercise.



Since $Y = X - X'$ has the same distribution as $-Y$ (that is, Y has a symmetric distribution), the trick of considering Y instead of X is called the **symmetrization** trick or symmetrization device. The symmetrization argument is useful in a variety of contexts, but may not give the best possible constants. In fact, in the next exercise, you are asked to sharpen the result of Exercise 5.12.

5.13 [Hoeffding's lemma] Suppose that X is zero-mean and $X \in [a, b]$ almost surely for constants $a < b$.

- (a) Show that X is $(b-a)/2$ -subgaussian.
- (b) Prove Hoeffding's inequality (5.8).



For part (a) it suffices to prove that $\psi_X(\lambda) \leq \lambda^2(b-a)^2/4$. By Taylor's theorem, for some λ' between 0 and λ , $\psi_X(\lambda) = \psi_X(0) + \psi'_X(0)\lambda + \psi''_X(\lambda')\lambda^2/2$. To bound the last term, introduce the distribution P_λ for $\lambda \in \mathbb{R}$ arbitrary: $P_\lambda(dz) = e^{-\psi_X(\lambda)}e^{\lambda z}P(dz)$. Show that $\Psi''_X(\lambda) = \mathbb{V}[Z]$ where $Z \sim P_\lambda$. Now, since $Z \in [a, b]$ with probability one, argue (without relying on $\mathbb{E}[Z]$) that $\mathbb{V}[Z] \leq (b-a)^2/4$.

5.14 Let X_p be a Bernoulli distribution with mean p , which means that $\mathbb{P}(X=1)=p$ and $\mathbb{P}(X=0)=1-p$.

- (a) Show that X_p is $1/2$ -subgaussian for all p .
- (b) Let $Q : [0, 1] \rightarrow [0, 1/2]$ be the function given by $Q(p) = \sqrt{\frac{1-2p}{2\ln((1-p)/p)}}$ where undefined points are defined in terms of their limits. Show that X_p is $Q(p)$ -subgaussian.
- (c) Plot $Q(p)$ as a function of p . How does it compare to $\sqrt{\mathbb{V}[X_p]} = \sqrt{p(1-p)}$?



Readers looking for a hint to part (b) in the previous exercise might like to look at the short paper by [Ostrovsky and Sirota \[2014\]](#).

5.15 In this question we try to understand the concentration of the empirical mean for Bernoulli random variables. Let X_1, X_2, \dots, X_n be independent Bernoulli random variables with mean $p \in [0, 1]$ and $\hat{p}_n = \sum_{t=1}^n X_t/n$. Let Z_n be normally distributed random variable with mean p and variance $p(1-p)/n$.

- (a) Write down expressions for $\mathbb{E}[\hat{p}_n]$ and $\mathbb{V}[\hat{p}_n]$.
- (b) What does the central limit theorem say about the relationship between \hat{p}_n and Z_n as n gets large?
- (c) For each $p \in \{1/10, 1/2\}$ and $\delta = 1/100$ and $\Delta = 1/10$ find the minimum n such that $\mathbb{P}(\hat{p}_n \geq p + \Delta) \leq \delta$.

(d) Let $p = 1/10$ and $\Delta = 1/10$ and

$$\begin{aligned} n_1(\delta, p, \Delta) &= \min \{n : \mathbb{P}(\hat{p}_n \geq p + \Delta) \leq \delta\} \\ n_2(\delta, p, \Delta) &= \min \{n : \mathbb{P}(Z_n \geq p + \Delta) \leq \delta\}. \end{aligned}$$

(i) Evaluate empirically or analytically the value of

$$\lim_{\delta \rightarrow 0} \frac{n_1(\delta, 1/10, 1/10)}{n_2(\delta, 1/10, 1/10)}$$

(ii) In light of the central limit theorem, explain why the answer you got in (i) was not 1.

5.16 Let X_1, \dots, X_n be a sequence of independent random variables with $X_t - \mathbb{E}[X_t] \leq b$ almost surely and $S_n = \sum_{t=1}^n (X_t - \mathbb{E}[X_t])$ and $V_n = \sum_{t=1}^n \mathbb{V}[X_t]$.

- (a) Show that $g(x) = \frac{1}{2} + \frac{x}{3!} + \frac{x^2}{4!} + \dots = (\exp(x) - 1 - x)/x^2$ is monotone increasing.
- (b) Let X be a random variable with $\mathbb{E}[X] = 0$ and $X \leq b$ almost surely. Show that $\mathbb{E}[\exp(X)] \leq 1 + g(b)\mathbb{V}[X]$.
- (c) Prove that $(1 + \alpha)\log(1 + \alpha) - \alpha \geq \frac{3\alpha^2}{6+2\alpha}$ for all $\alpha \geq 0$.
- (d) Let $\varepsilon > 0$ and $\alpha = b\varepsilon/V$ and prove that

$$\begin{aligned} \mathbb{P}\left(\sum_{t=1}^n (X_t - \mathbb{E}[X_t]) \geq \varepsilon\right) &\leq \exp\left(-\frac{V_n}{b^2} ((1 + \alpha)\log(1 + \alpha) - \alpha)\right) \\ &\leq \exp\left(-\frac{\varepsilon^2}{2V(1 + \frac{b\varepsilon}{3V})}\right). \end{aligned} \tag{5.10}$$

(e) Use the previous result to show that

$$\mathbb{P}\left(S_n \geq \sqrt{2 \sum_{t=1}^n \mathbb{V}[X_t] \log\left(\frac{1}{\delta}\right)} + \frac{2b}{3n} \log\left(\frac{1}{\delta}\right)\right) \leq \delta.$$

(f) What can be said if X_1, \dots, X_n are Gaussian? Discuss empirically or theoretically whether or not a dependence on b is avoidable or not.



The bound in Eq. (5.10) is called Bernstein's inequality. There are several generalizations, the most notable of which is the martingale version that slightly relaxes the independence assumption. We will see martingale techniques in Chapter 20. Another useful variant (under slightly different conditions) replaces the actual variance with the empirical variance. This is useful in the common case that the variance is unknown. For more on this see papers by Audibert et al. [2007], Mnih et al. [2008], Maurer and Pontil [2009] or skip ahead to Exercise 7.7.

5.17 Let X_1, X_2, \dots, X_n be a sequence of random variables adapted to filtration $\mathbb{F} = (\mathcal{F}_t)_t$. Abbreviate $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_t]$ and $\mu_t = \mathbb{E}_{t-1}[X_t]$. Suppose that $\eta > 0$ satisfies $\eta X_t \leq 1$ almost surely. Prove that

$$\mathbb{P}\left(\sum_{t=1}^n (X_t - \mu_t) \geq \eta \sum_{t=1}^n \mathbb{E}_{t-1}[X_t^2] + \frac{1}{\eta} \log\left(\frac{1}{\delta}\right)\right) \leq \delta.$$

 Use Chernoff's method and the fact that $\exp(x) \leq 1 + x + x^2$ for all $x \leq 1$ and $\exp(x) \geq 1 + x$ for all x .

5.18 Let X_1, \dots, X_n be independent random variables with $\mathbb{P}(X_t \leq x) \leq x$ for each $x \in [0, 1]$ and $t \in [n]$. Prove for any $\varepsilon > 0$ that

$$\mathbb{P}\left(\sum_{t=1}^n \log(1/X_t) \geq \varepsilon\right) \leq \left(\frac{\varepsilon}{n}\right)^n \exp(n - \varepsilon).$$

5.19 Let X_1, \dots, X_n be an independent and identically distributed sequence taking values in $[m]$. For $i \in [m]$ let $p(i) = \mathbb{P}(X_1 = i)$ and $\hat{p}(i) = \frac{1}{n} \sum_{t=1}^n \mathbb{I}\{X_t = i\}$. Show that for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(\|p - \hat{p}\|_1 \geq \sqrt{\frac{2m \log(2/\delta)}{n}}\right) \leq \delta. \quad (5.11)$$

 This is quite a tricky exercise. The result is due to Weissman et al. [2003]. It is worth comparing this to what can be obtained from Hoeffding's inequality, which implies for any $i \in [m]$ and $\delta \in (0, 1)$ that with probability $1 - \delta$,

$$|\hat{p}(i) - p(i)| < \sqrt{\frac{2 \log(2/\delta)}{n}}.$$

By a union bound this ensures that with probability $1 - \delta$,

$$\sum_i |\hat{p}(i) - p(i)| < m \sqrt{\frac{2 \log(2m/\delta)}{n}},$$

which is significantly weaker than the upper bound in (5.11). A standard approach to deriving a stronger inequality is to use the fact that $\|p - \hat{p}\|_1 = \sup_{x: \|x\|_\infty \leq 1} \langle p - \hat{p}, x \rangle$. Choose finite subset $S \subset B = [-1, 1]^m$ such that for any point in $x \in B$ there exists a $y \in S$ such that $\|x - y\|_\infty \leq \varepsilon/3$. Let $x^* = \operatorname{argmax}_{x \in B} \langle p - \hat{p}, x \rangle$ and $s = \operatorname{argmin}_{u \in S} \|s - x\|_\infty$. Then $\langle p - \hat{p}, x^* \rangle = \langle p - \hat{p}, s \rangle + \langle p - \hat{p}, s - x^* \rangle \leq \langle p - \hat{p}, s \rangle + 2 \sup_{p: \|p\|_1 \leq 1} \langle p, s - x^* \rangle = \langle p - \hat{p}, s \rangle + 2\|s - x^*\|_\infty = \langle p - \hat{p}, s \rangle + 2\varepsilon/3$. By applying Hoeffding's inequality to $\langle p - \hat{p}, s \rangle$ and a union bound we see that if $n \geq 2 \log(|S|/\delta)/\varepsilon^2$, then with probability $1 - \delta$ it holds that $\|p - \hat{p}\|_1 \leq \varepsilon$. Choosing S to have the fewest elements gives a bound similar to that of Lemma 37.2. In particular, S can be

chosen to be the regular grid with stride $\varepsilon/3$, giving $|S| = (6/\varepsilon)^m$. The quantity $\sup_{x \in X} \langle p - \hat{p}, x \rangle$ is called an **empirical process**. Such empirical processes are the subject of extensive study in the field of empirical process theory, which has many applications within statistics, machine learning and also beyond these field in almost all areas of mathematics [Vaart and Wellner, 1996, Dudley, 2014, van de Geer, 2000].

5.20 Let X_1, X_2, \dots, X_n be a sequence of nonnegative random variables adapted to filtration $(\mathcal{F}_t)_{t=0}^n$ such that $\sum_{t=1}^n X_t \leq 1$ almost surely. Prove that for all $x > 1$,

$$\mathbb{P}\left(\sum_{t=1}^n \mathbb{E}[X_t | \mathcal{F}_{t-1}] \geq x\right) \leq f_n(x) = \begin{cases} \left(\frac{n-x}{n-1}\right)^{n-1} & \text{if } x < n \\ 0 & \text{if } x \geq n, \end{cases}$$

where the equality serves as the definition of $f_n(x)$.



This problem does not use the techniques introduced in the chapter. Prove that Bernoulli random variables are the worst case and use backwards induction. Although this result is new to our knowledge, a weaker version was derived by Kirschner and Krause [2018] for the analysis of information directed sampling. The bound is tight in the sense that there exists a sequence of random variables and filtration for which equality holds.

Part II

Stochastic Bandits with Finitely Many Arms

In the first part devoted entirely to bandits we introduce the fundamental algorithms and many ideas of analysis for unstructured finite-armed bandits. The keywords here are finite, unstructured and stochastic. The first of these just means that the number of actions available is finite. The second is more ambiguous, but roughly means that choosing one action yields no information about the mean payoff of other arms. Finally, the ‘stochastic’ keyword means that the sequence of rewards associated with each action is independent and identically distributed according to some distribution. This latter assumption will be relaxed in Part III.

There are several reasons to study this class of bandit problems. For one thing, their simplicity makes them relatively easy to analyze and permits a deep understanding of the tradeoff between exploration and exploitation. Secondly, many of the algorithms designed for finite-armed bandits, and the principle underlying them, can be generalized to other settings. Finally, finite-armed bandits already have applications. Notably as a replacement to A/B testing as discussed in the introduction.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

6 The Explore-then-Commit Algorithm

With the background on concentration out of the way, we are ready to present the first bandit policy of the book. The policy we present is one of the simplest one that one can imagine: It says to first explore by choosing each arm a certain number of times and subsequently exploit by playing the arm that appeared best after exploration.



For this chapter, as well as Chapters 7 to 9, we assume that all bandit instances are in $\mathcal{E}_{\text{SG}}^K(1)$, which means the reward distribution for all arms is 1-subgaussian.

The focus on subgaussian distributions is mainly for simplicity. Many of the techniques in the chapters that follow can be applied to other stochastic bandits such as those listed in Table 4.1. The key difference is that new concentration analysis is required that exploits the different assumptions. The Bernoulli case is covered in Chapter 10 where other cases are briefly discussed along with references to the literature. As this stage, what is important to keep in mind is that larger environment classes correspond to harder bandit problems and we expect to see the effect of this in the strength of results.

Notice that the subgaussian assumption restricts the subgaussian constant to $\sigma = 1$, which saves us from endlessly writing σ . All results hold for other subgaussian constants by scaling the rewards (see Lemma 5.2). Two points are obscured by this simplification.

- 1 All the algorithms that follow rely on the knowledge of σ .
- 2 It may happen sometimes that P_i is subgaussian for all arms, but with a different subgaussian constant for each arm. Algorithms are easily adapted to this situation if the subgaussian constants are known, a process we leave to the readers in Exercise 7.2. The situation where σ^2 is not known is more complicated and is discussed in Chapter 7 (and see Exercise 7.8).

We now describe the **explore-then-commit** (ETC) policy, which is characterized by the number of times it explores each arm, denoted by a natural number m . Because there are K actions, the algorithm will explore for mK rounds before choosing a single action for the remaining rounds. In order to define this policy formally we need a little more notation. Let $\hat{\mu}_i(t)$ be the average

pay-off received from arm i up to round t . Hence

$$\hat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^t \mathbb{I}\{A_s = i\} X_s,$$

where recall that $T_i(t) = \sum_{s=1}^t \mathbb{I}\{A_s = i\}$ is the number of times action i is chosen up to the end of round t . The vanilla version of an explore-then-commit policy is given below. Recall for natural numbers $a, b \geq 1$ that $a \bmod b$ is the remainder when a is divided by b . Or equivalently, $a \bmod b = a - b \lfloor a/b \rfloor$.

1: **Input** $m \in \mathbb{N}$.

2: In round t choose action

$$A_t = \begin{cases} i, & \text{if } (t \bmod K) + 1 = i \text{ and } t \leq mK; \\ \operatorname{argmax}_i \hat{\mu}_i(mK), & t > mK. \end{cases}$$

(ties in the argmax are broken arbitrarily)

Algorithm 1: Explore-then-commit policy

The formal definition of the explore-then-commit policy leads rather immediately to the analysis of its regret.

THEOREM 6.1 (Regret of ETC) *The the expected regret of the ETC policy is bounded by,*

$$R_n \leq \left(m \wedge \left\lceil \frac{n}{K} \right\rceil \right) \sum_{i=1}^K \Delta_i + (n - mK)^+ \sum_{i=1}^K \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right).$$

Proof By the decomposition given in Lemma 4.2 the regret can be written as

$$R_n = \sum_{i=1}^K \Delta_i \mathbb{E}[T_i(n)].$$

In the first mK rounds the ETC policy is completely deterministic, choosing each action exactly m times. Subsequently it chooses a single action that gave the largest average payoff during exploration. Thus,

$$\begin{aligned} \mathbb{E}[T_i(n)] &\leq m \wedge \left\lceil \frac{n}{K} \right\rceil + (n - mK)^+ \mathbb{P}(i = A_{mK+1}) \\ &\leq m \wedge \left\lceil \frac{n}{K} \right\rceil + (n - mK)^+ \mathbb{P}\left(\hat{\mu}_i(mK) \geq \max_{j \neq i} \hat{\mu}_j(mK)\right), \end{aligned}$$

where recall that $a \wedge b = \min(a, b)$ and $(x)^+ = \max(x, 0)$ denotes the positive part of $x \in \mathbb{R}$ (thus, the second term is zero when $n < mK$). Now we need to bound the probability in the second term above. For the sake of simplifying the presentation, assume without loss of generality that the arm one is an optimal

arm so that $\mu_1 = \mu^* = \max_i \mu_i$ (though the learner does not know this). Then

$$\begin{aligned} \mathbb{P}\left(\hat{\mu}_i(mK) \geq \max_{j \neq i} \hat{\mu}_j(mK)\right) &\leq \mathbb{P}(\hat{\mu}_i(mK) \geq \hat{\mu}_1(mK)) \\ &= \mathbb{P}(\hat{\mu}_i(mK) - \mu_i - (\hat{\mu}_1(mK) - \mu_1) \geq \Delta_i). \end{aligned}$$

The next step is to check that $\hat{\mu}_i(mK) - \mu_i - (\hat{\mu}_1(mK) - \mu_1)$ is $\sqrt{2/m}$ -subgaussian, which by the properties of subgaussian random variables follows from the definitions of $\{\hat{\mu}_j\}_j$ and the algorithm. Therefore, by Corollary 5.1 we have

$$\mathbb{P}(\hat{\mu}_i(mK) - \mu_i - \hat{\mu}_1(mK) + \mu_1 \geq \Delta_i) \leq \exp\left(-\frac{m\Delta_i^2}{4}\right)$$

and by straightforward substitution we obtain

$$R_n \leq \left(m \wedge \left\lceil \frac{n}{K} \right\rceil\right) \sum_{i=1}^K \Delta_i + (n - mK)^+ \sum_{i=1}^K \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right). \quad (6.1)$$

□

Although we will discuss many disadvantages of the approach taken by ETC, the above bound cleanly illustrates the fundamental challenge faced by the learner, which is the trade-off between exploration and exploitation. If m is large, then the policy explores for too long and the first term will be eventually too large. On the other hand, if m is too small, then the probability that the algorithm commits to the wrong arm will grow and the second term becomes too large. The question is how to choose m ? If we limit ourselves to $K = 2$ and if the first arm is optimal then $\Delta_1 = 0$ and by using $\Delta \doteq \Delta_2$ the above display simplifies to

$$R_n \leq m\Delta + (n - 2m)^+ \Delta \exp\left(-\frac{m\Delta^2}{4}\right) \leq m\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right). \quad (6.2)$$

Provided that n is reasonably large the quantity on the right-hand side of the above display is minimized (up to a possible rounding error) by

$$m = \max \left\{ 0, \left\lceil \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right) \right\rceil \right\} \quad (6.3)$$

and for this choice (and any n) the regret is bounded by

$$R_n \leq \Delta + \frac{4}{\Delta} \left(1 + \max \left\{ 0, \log\left(\frac{n\Delta^2}{4}\right) \right\} \right). \quad (6.4)$$

Notice that here Δ appears in the denominator of the regret bound, which means that as Δ becomes very small the regret bound grows unboundedly. Is that reasonable and why is it happening? The explanation is simply that in the second inequality of (6.2) we have upper bounded $(n - mK)^+$ by n also $m \wedge \lceil n/K \rceil$ by m , regardless of the value of n , whereas (using that $K = 2$) for $n < 2m$, $(n - 2m)^+ = 0$ and $m \wedge \lceil n/2 \rceil \leq \lceil n/2 \rceil$. In fact, for any n (and regardless of what

policy is being used!) $R_n = \Delta \mathbb{E}[T_2(n)] \leq \Delta n$. Taking the minimum of this and the bound shown in (6.4), we get

$$R_n \leq \min \left\{ n\Delta, \Delta + \frac{4}{\Delta} \left(1 + \log \left(\frac{n\Delta^2}{4} \right) \right) \right\}. \quad (6.5)$$

We leave it as an exercise to the reader to check that $R_n = O(\sqrt{n})$ regardless of the value of Δ . Bounds of this nature are usually called **worst-case** or **problem-free** or **problem independent** (cf. Eq. (4.2) or Eq. (4.3)). The reason is that the regret depends only on the distributional assumption (i.e., the choice of the environment \mathcal{E}^K), but not on the specific distribution $\nu \in \mathcal{E}^K$. Sometimes bounds of this type are also called gap-free as they do not depend on Δ . In contrast, bounds that depend on the sub-optimality gaps are called **gap/problem/distribution dependent**.

Later you will see that the bound in (6.4) is very close to optimal in a number of ways. However, there is one big caveat. The choice of m given above (which defines the policy) depends on both the sub-optimality gap and the horizon. While sometimes the horizon might be known in advance, it is practically never reasonable to assume knowledge of the sub-optimality gaps. So is there a reasonable way to choose m that does not depend on the unknown gap, but may depend on n ? It turns out that there is a choice that will achieve $R_n = O(n^{2/3})$ regardless of the value of Δ , i.e., in a worst-case sense (to see this, recall that $R_n \leq \Delta n$). The need to know the sub-optimality gap can be overcome by allowing m to be data-dependent. That is, the learner chooses each arm alternately until it decides based on its observations to commit to a single arm for the remaining rounds. We return to this point briefly in the notes at the end of the section, but will not cover the details because it turns out there are other algorithms that are superior in practice and do not even need to know the horizon.



So how does this play out on actual data? To examine this question we plot the expected regret of ETC when playing a two-armed bandit with means $\mu_1 = 0$ and $\mu_2 = -\Delta$. The horizon is set to $n = 1000$ and $\Delta > 0$ is varied between 0 and 1. The plot shows three curves:

- (a) The theoretical upper bound given in Eq. (6.5).
- (b) The regret of the ETC algorithm with m set as suggested in Eq. (6.3).
- (c) The regret of the ETC algorithm with the optimal m , which may be calculated numerically using the assumption that the noise is exactly Gaussian.

Each data point is the average of 10^4 simulations, which means that error-bars are so small that they are not visible (and hence no attempt is made to show them).

The figure shows the actual performance of the ETC algorithm roughly tracks the theory, and the optimally tuned version provides a modest improvement in performance. It is important to emphasize that the optimally tuned algorithm

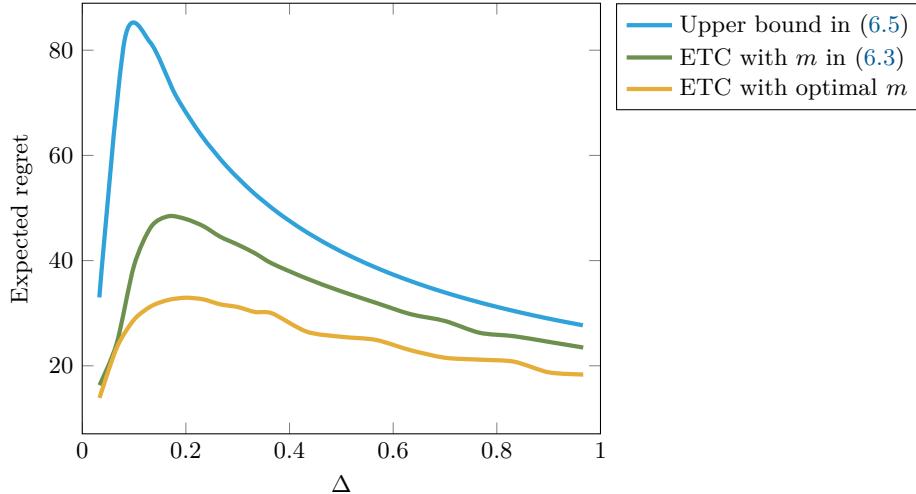


Figure 6.1 Expected regret of ETC strategies

is ‘cheating’ in an even stronger way than the choice based on (6.3) because the calculation of the optimal m was based on a Gaussian assumption, while the choice given in Eq. (6.3) only relied on the noise being subgaussian.

6.1 Notes

- 1 An algorithm is called **anytime** if it does not require advance knowledge of the horizon n . As discussed, the ETC algorithm is not anytime because the choice of commitment time depends on the horizon. This limitation can be addressed in a general way by using the **doubling trick**, which is a simple way to convert a horizon-dependent algorithm into an **anytime algorithm** that does not depend on the horizon. We explain in Exercise 6.6.
- 2 By allowing the exploration time m to be a data-dependent random variable it is possible to recover near-optimal regret without knowing the sub-optimality gaps. For examples where this is done, see the articles by [Auer and Ortner \[2010\]](#) and [Garivier et al. \[2016b\]](#) and Exercise 6.5.

6.2 Bibliographical remarks

Explore-then-commit has a long history. [Robbins \[1952\]](#) considered ‘certainty equivalence with forcing’, which chooses the arm with the largest sample mean except at a fixed set of times $T_k \subset \mathbb{N}$ when arm k is chosen for $k \in [K]$. By

choosing the set of times carefully, it is shown that this policy enjoys sublinear regret. While ETC performs all the exploration at the beginning, Robbins's policy spreads the exploration over time. This is advantageous if the horizon is not known, but is disadvantageous when the horizon is known (why would want to delay exploration when the horizon is known?). Anscombe [1963] considered exploration and commitment in the context of medical trials or other experimental setups. He already largely solves the problem in the Gaussian case and highlights many of the important considerations. Besides this, the article is beautifully written and well worth reading. Strategies based on exploration and commitment are simple to implement and analyze. They can also generalize well to more complex settings. For example, Langford and Zhang [2008] considers this style of policy under the name ‘epoch-greedy’ for contextual bandits (the idea of exploring then exploiting in epochs, or intervals, is essentially what Robbins [1952] suggested). We'll return to contextual bandits in Chapter 18. Abbasi-Yadkori et al. [2009] (see also Abbasi-Yadkori 2009b) and Rusmevichientong and Tsitsiklis [2010] consider ETC-style policies under the respective names of ‘forced exploration’ and ‘phased exploration and greedy exploitation’ (PEGE) in the context of linear bandits (which we shall meet in Chapter 19). Other names include ‘forced sampling’, ‘explore-first’, ‘explore-then-exploit’. Garivier et al. [2016b] have shown that ETC policies are necessarily suboptimal in the limit of infinite data in a way that is made precise in Chapter 16. As mentioned earlier, ε -greedy is a close relative of ETC and can be viewed as the randomized version of Robbin's algorithm. One may ask why would want to randomize? The best answer is simplicity: The policy only depends on the exploration parameter ε rather than some complicated schedule. This can be useful in complicated settings like reinforcement learning where many instances of the same algorithm are acting simultaneously and communication is costly. In Chapter 11 we'll see the role of randomization when the bandit itself is allowed to react to the actions of the learner in a malicious way. The history of ε -greedy is unclear, but it is a popular and widely used and known algorithm in reinforcement learning [Sutton and Barto, 1998]. Auer et al. [2002a] analyze the regret of ε -greedy with slowly decreasing ε (see Exercise 6.7). There are other kinds of randomized exploration as well, including Thompson sampling [1933] and Boltzmann exploration analyzed recently by Cesa-Bianchi et al. [2017].

6.3 Exercises

6.1 In the proof of Theorem 6.1 we wrote: “The next step is to check that $\hat{\mu}_i(mK) - \mu_i - (\hat{\mu}_1(mK) - \mu_1)$ is $\sqrt{2/m}$ -subgaussian, which by the properties of subgaussian random variables follows from the definitions of $\{\hat{\mu}_j\}_j$ and the algorithm.” Prove this in a fully rigorous manner. You can use the interaction protocol, the assumption that rewards are 1-subgaussian, the definition of $\{\hat{\mu}_j\}_j$ and the definition of ETC. In particular, prove that $\hat{\mu}_j(mK)$ is the sample mean

of m i.i.d. random variables chosen from P_j . Note that this is *not* the definition of $\hat{\mu}_j(mK)$. Further, the interaction protocol only specifies that $X_t \sim P_{A_t}$, independently of $(A_1, X_1, \dots, A_{t-1}, X_{t-1})$ given A_t .

6.2 Fix $\delta \in (0, 1)$. Modify the ETC algorithm to depend on δ and prove a bound on the pseudo-regret $\bar{R}_n = n\mu^* - \sum_{t=1}^n \mu_{A_t}$ of ETC that holds with probability $1 - \delta$.

6.3 Fix $\delta \in (0, 1)$. Prove a bound on the random regret $\hat{R}_n = n\mu^* - \sum_{t=1}^n X_t$ of ETC that holds with probability $1 - \delta$. Compare this to the bound derived for the pseudo-regret in the previous exercise. What can you conclude?

6.4 In this exercise we investigate the empirical behavior of the Explore-Then-Commit algorithm on a two-armed Gaussian bandit with means $\mu_1 = 0$ and $\mu_2 = -\Delta$. Let

$$\bar{R}_n = \sum_{t=1}^n \Delta_{A_t},$$

which is chosen so that $R_n = \mathbb{E}[\bar{R}_n]$. Complete the following:

- (a) Using programming language of your choice, write a function that accepts an integer n and $\Delta > 0$ and returns the value of m that *exactly* minimizes the expected regret.
- (b) Reproduce Fig. 6.1, which shows the expected regret of the ETC algorithm for different choices of m as a function of Δ .
- (c) Now fix $\Delta = 1/10$ and plot the expected regret as a function of m with $n = 2000$. Your plot should resemble Fig. 6.2.
- (d) Plot the variance $\mathbb{V}[\bar{R}_n]$ as a function of m for the same bandit as above. Your plot should resemble Fig. 6.3.
- (e) Explain the shape of the curves you observed in Parts (b), (c) and (d) and reconcile what you see with the theoretical results.
- (f) Think, experiment and plot. Is it justified to plot $\mathbb{V}[\bar{R}_n]$ as a summary of how \bar{R}_n is distributed? Explain your thinking.

6.5 In this question we investigate how far we can push the ETC algorithm. Assume for the purpose of this exercise that ETC interacts with a 2-armed 1-subgaussian bandit with means $\mu_1, \mu_2 \in \mathbb{R}$ such that $\Delta = |\mu_1 - \mu_2|$.

- (a) Find a choice of m that depends only on the horizon n and *not* Δ such the regret of Algorithm 1 is bounded by

$$R_n \leq C(\Delta + n^{2/3}),$$

where $C > 0$ is a universal constant.

- (b) Now suppose the commitment time is allowed to be data-dependent, which means the algorithm explores each arm alternately until some condition is

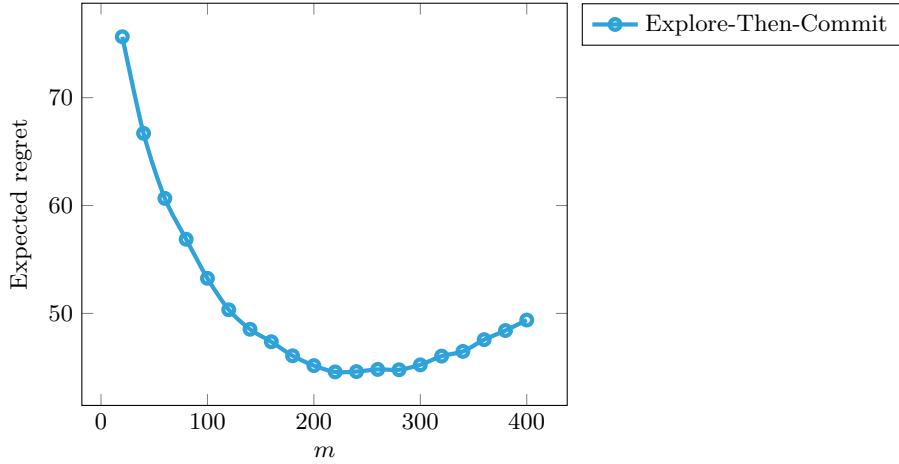


Figure 6.2 Expected regret for Explore-Then-Commit over 10^5 trials on a Gaussian bandit with means $\mu_1 = 0, \mu_2 = -1/10$

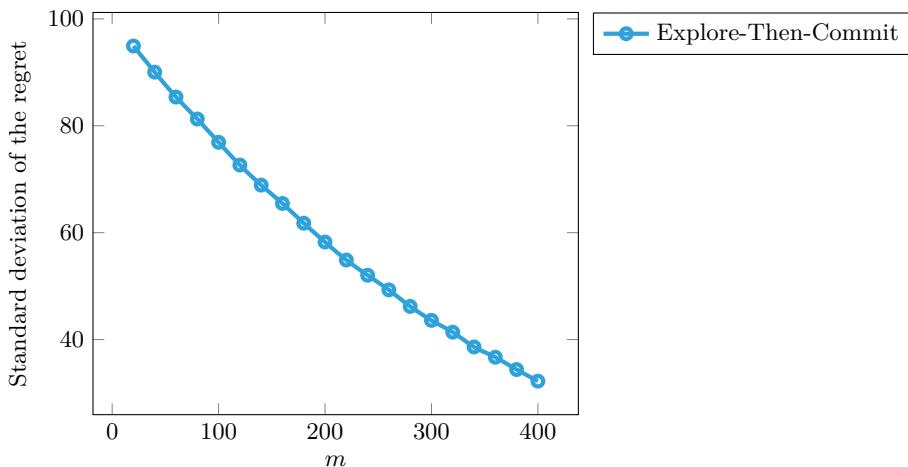


Figure 6.3 Standard deviation of the regret for ETC over 10^5 trials on a Gaussian bandit with means $\mu_1 = 0, \mu_2 = -1/10$

met and then commits to a single arm for the remainder. Design a condition such that the regret of the resulting algorithm can be bounded by

$$R_n \leq C \left(\Delta + \frac{\log n}{\Delta} \right), \quad (6.6)$$

where C is a universal constant. Your condition should only depend on the observed rewards and the time horizon. It should *not* depend on μ_1, μ_2 or Δ .

- (c) Show that any algorithm for which (6.6) holds also satisfies $R_n \leq C(\Delta + \sqrt{n \log(n)})$ for suitably chosen universal constant $C > 0$.

- (d) As for part (b), but now the objective is to design a condition such that the regret of the resulting algorithm is bounded by

$$R_n \leq C \left(\Delta + \frac{\log \max \{e, n\Delta^2\}}{\Delta} \right), \quad (6.7)$$

- (e) Show that any algorithm for which (6.7) holds also satisfies $R_n \leq C(\Delta + \sqrt{n})$ for suitably chosen universal constant $C > 0$.



For Part (a) start from $R_n \leq m\Delta + n\Delta \exp(-m\Delta^2/2)$ and assume that the second term here dominates the first term. Find Δ maximizing the resulting regret upper bound. Based on this propose m and verify that the starting assumption has been met by your choice regardless the values of n and Δ . For Part (b) think about the simplest stopping policy and then make it ‘robust’ by using confidence intervals. Tune the failure probability. For Part (c) note that the regret can never be larger than $n\Delta$.



In the later parts of Exercise 6.5 we allowed the commitment time to be data-dependent. This makes it a random variable and so it should be capitalized to M instead of m . In the language of formal probability, the random variable $T = 2M$ is called a stopping time with respect to the filtration $(\mathcal{F}_t)_{t \geq 1}$ where $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$. We’ll introduce the formal definition of a stopping time in a later chapter, but for now we mention briefly that it means that the event $\{T = t\}$ is \mathcal{F}_t -measurable for each t . The curious reader might like to think about what this definition means. (**Hint:** The algorithm cannot commit on the basis of information it does not have!)

- 6.6** Let \mathcal{E} be an arbitrary class of bandits (for example, $\mathcal{E} = \mathcal{E}_{SG}^K$). Suppose you are given a policy \mathcal{A} designed for \mathcal{E} that accepts the horizon n as a parameter and has a regret guarantee of

$$R_n \leq f_n(\nu), \quad \forall \nu \in \mathcal{E},$$

where $f_n : \mathcal{E} \rightarrow [0, \infty)$ is a sequence of functions. The purpose of this exercise is to analyze a meta-algorithm based on the so-called **doubling trick** that converts a policy depending on the horizon to a policy with similar guarantees that does not. Let $n_1 < n_2 < n_3 < \dots$ be a fixed sequence of integers and consider the policy that runs \mathcal{A} with horizon n_1 until round $t = \max\{n, n_1\}$. Then restarts the algorithm with horizon n_2 until $t = \max\{n, n_1 + n_2\}$. Then restarts again with horizon n_3 until $t = \max\{n, n_1 + n_2 + n_3\}$ and so-on.

- (a) Let $\ell_{\max} = \min\{\ell : \sum_{i=1}^{\ell} n_i \geq n\}$. Prove that the regret of the meta-algorithm

is at most

$$R_n \leq \sum_{\ell=1}^{\ell_{\max}} f_{n_\ell}(\nu).$$

- (b) Suppose that $f_n(\nu) \leq \sqrt{n}$. Show that if $n_\ell = 2^{\ell-1}$, then the regret of the meta-algorithm is at most

$$R_n \leq C\sqrt{n},$$

where $C > 0$ is a carefully chosen universal constant.

- (c) Suppose that $f_n(\nu) = g(\nu) \log(n)$ for some function $g : \mathcal{E} \rightarrow [0, \infty)$. What is the regret of the meta-algorithm if $n_\ell = 2^{\ell-1}$? Can you find a better choice of $(n_\ell)_\ell$?
- (d) In light of this idea, should we bother trying to design algorithms that do not depend on the horizon? Are there any disadvantages to using the doubling trick? If so, what are they?

6.7 For this exercise assume the rewards are 1-subgaussian and there are $K \geq 2$ arms. The ε -greedy algorithm depends on a sequence of parameters $\varepsilon_1, \varepsilon_2, \dots$. First it chooses each arm once and subsequently chooses $A_t = \operatorname{argmax}_i \hat{\mu}_i(t-1)$ with probability $1 - \varepsilon_t$ and otherwise chooses an arm uniformly at random.

- (a) Prove that if $\varepsilon_t = \varepsilon > 0$, then $\lim_{n \rightarrow \infty} \frac{R_n}{n} = \frac{\varepsilon}{K} \sum_{i=1}^K \Delta_i$.
- (b) Let $\Delta_{\min} = \min \{\Delta_i : \Delta_i > 0\}$ and let $\varepsilon_t = \min \left\{ 1, \frac{CK}{t\Delta_{\min}^2} \right\}$ where $C > 0$ is a sufficiently large universal constant. Prove that there exists a universal $C' > 0$ such that

$$R_n \leq C' \sum_{i=1}^K \left(\Delta_i + \frac{\Delta_i}{\Delta_{\min}^2} \log \max \left\{ e, \frac{n\Delta_{\min}^2}{K} \right\} \right).$$

6.8 A simple way to generalize the ETC policy to multiple arms and overcome the problem of tuning the commitment time is to use an **elimination algorithm**. The algorithm operates in phases and maintains an **active set** of arms that could be optimal. In the ℓ th phase the algorithm aims to eliminate from the active set all arms i for which $\Delta_i \geq 2^{-\ell}$.

Without loss of generality, assume that arm 1 is an optimal arm.

- (a) Show that for any $\ell \geq 1$,

$$\mathbb{P}(1 \notin A_{\ell+1}, 1 \in A_\ell) \leq K \exp \left(-\frac{\tau_\ell 2^{-2\ell}}{4} \right)$$

- (b) Show that if $i \in [K]$ and $\ell \geq 1$ are such that $\Delta_i \geq 2^{-\ell}$, then

$$\mathbb{P}(i \in A_{\ell+1}, 1 \in A_\ell, i \in A_\ell) \leq \exp \left(-\frac{\tau_\ell (\Delta_i - 2^{-\ell})^2}{4} \right)$$

```

1: Input:  $K$  and sequences  $(\tau_\ell)_\ell$ 
2:  $A_1 = \{1, 2, \dots, K\}$ 
3: for  $\ell = 1, 2, 3, \dots$  do
4:   Choose  $i \in A_\ell$  each arm  $\tau_\ell$  times
5:   Let  $\hat{\mu}_{i,\ell}$  be the average reward for arm  $i$  from this phase
6:   Update active set:

```

$$A_{\ell+1} = \left\{ i : \hat{\mu}_{i,\ell} + 2^{-\ell} \geq \max_{j \in A_\ell} \hat{\mu}_{j,\ell} \right\}$$

```
7: end for
```

- (c) Let $\ell_i = \min \{\ell \geq 1 : 2^{-\ell} \leq \Delta_i/2\}$. Choose τ_ℓ in such a way that $\mathbb{P}(\exists \ell : 1 \notin A_\ell) \leq 1/n$ and $\mathbb{P}(i \in A_{\ell_i+1}) \leq 1/n$.

- (d) Show that your algorithm has regret at most

$$R_n \leq C \sum_{i:\Delta_i > 0} \left(\Delta_i + \frac{1}{\Delta_i} \log(n) \right),$$

where $C > 0$ is a carefully chosen universal constant.

- (e) Modify your choice of τ_ℓ (if necessary) to derive a regret bound

$$R_n \leq C \sum_{i:\Delta_i > 0} \left(\Delta_i + \frac{1}{\Delta_i} \log \max \{e, n\Delta_i^2\} \right).$$

- 6.9** For this exercise assume the rewards are 1-subgaussian. The ε -greedy algorithm depends on a sequence of parameters $\varepsilon_1, \varepsilon_2, \dots$. First it chooses each arm once and subsequently chooses $A_t = \operatorname{argmax}_i \hat{\mu}_i(t-1)$ with probability $1 - \varepsilon_t$ and otherwise an arm uniformly at random.

- (a) Prove that if $\varepsilon_t = \varepsilon > 0$, then $\lim_{n \rightarrow \infty} \frac{R_n}{n} = \frac{\varepsilon}{K} \sum_{i=1}^K \Delta_i$.
- (b) Let $\Delta_{\min} = \min \{\Delta_i : \Delta_i > 0\}$ and let $\varepsilon_t = \min \left\{ 1, \frac{CK}{t\Delta_{\min}^2} \right\}$ where $C > 0$ is a sufficiently large universal constant. Prove that there exists a universal $C' > 0$ such that

$$R_n \leq C' \sum_{i=1}^K \left(\Delta_i + \frac{\Delta_i}{\Delta_{\min}^2} \log \max \left\{ e, \frac{n\Delta_{\min}^2}{K} \right\} \right).$$

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

7 The Upper Confidence Bound Algorithm

We now describe the celebrated Upper Confidence Bound (UCB) algorithm, which offers several advantages over the ETC algorithm introduced in the last chapter:

- (a) It does not depend on advance knowledge of the suboptimality gaps.
- (b) It behaves well when there are more than two arms.
- (c) The version introduced here *does* depend on the horizon n , but in the next chapter we will see how to eliminate that as well.

The algorithm has many different forms, depending on the distributional assumptions on the noise. Like in the previous chapter, we assume the noise is 1-subgaussian. A serious discussion of other options is delayed until Chapter 10. The algorithm is based on the principle of **optimism in the face of uncertainty** (OFU), which is applicable to various exploration problems not only finite-armed stochastic bandits:



The optimism in the face of uncertainty principle states that one should choose their actions as if the environment is as nice as **plausibly possible**.

To illustrate the intuition, imagine visiting a new country and making a choice between sampling the local cuisine or visiting a well-known multinational chain. Taking an optimistic view of the unknown local cuisine leads to exploration because without data it *could* be amazing. Then, after trying the new option a few times you can update your statistics about each choice and make a more informed decision. On the other hand, taking a pessimistic view of the new option discourages exploration and you may suffer significant regret if the local options are delicious. Just how optimistic you should be is a difficult decision, which we explore for the rest of the chapter in the context of finite-armed bandits.

For bandits the optimism principle means using the data observed so far to assign to each arm a value called the **upper confidence bound** that with high probability is an overestimate of the unknown mean. The intuitive reason why this leads to sublinear regret is simple. Assuming the upper confidence bound assigned to the optimal arm is indeed an overestimate, then another arm can only be played if its upper confidence bound is larger than that of the optimal arm, which in turn is larger than the mean of the optimal arm. And yet this cannot happen too often because the additional data provided by playing a suboptimal

arm means that the upper confidence bound for this arm will eventually fall below that of the optimal arm.

This explains why the optimism principle will eventually get things right (that is, why it leads to sublinear regret). But the argument does not explain why an optimistic algorithm for finite-armed bandits should be as good (or better) than a well-tuned ETC. Whether this holds or not hinges on the exact definition of ‘plausible’. Recall that if X_1, X_2, \dots, X_n are independent and 1-subgaussian with mean μ and $\hat{\mu} = \sum_{t=1}^n X_t/n$, then by Eq. (5.6), for any $\delta \in [0, 1]$,

$$\mathbb{P}\left(\mu \geq \hat{\mu} + \sqrt{\frac{2 \log(1/\delta)}{n}}\right) \leq \delta. \quad (7.1)$$

When considering its options in round t the learner has observed $T_i(t-1)$ samples from arm i and received rewards from that arm with an empirical mean of $\hat{\mu}_i(t-1)$. Then a reasonable candidate for ‘as large as plausibly possible’ for the unknown mean of the i th arm is

$$\text{UCB}_i(t-1, \delta) \doteq \hat{\mu}_i(t-1) + \sqrt{\frac{2 \log(1/\delta)}{T_i(t-1)}}. \quad (7.2)$$

Great care is required when comparing (7.1) and (7.2) because in the former the number of samples is the constant n , but in the latter it is a *random variable* $T_i(t-1)$. By and large, however, this is merely an annoying technicality and the intuition remains that δ is approximately an upper bound on the probability of the event that the above quantity is an underestimate of the true mean. More details are given in Exercise 7.1.

At last we have everything we need to state a version of the UCB algorithm, which takes as input the number of arms and the error probability δ .

- 1: **Input** K and δ
- 2: Choose each action once
- 3: For rounds $t > K$ choose action

$$A_t = \operatorname{argmax}_i \text{UCB}_i(t-1, \delta)$$

Algorithm 2: UCB(δ) algorithm



Although there are many versions of the UCB algorithm, we often do not distinguish them by name and hope the context is clear. For the rest of this chapter we’ll usually call UCB(δ) just UCB.

The algorithm first chooses each arm once, which is necessary because the term inside the square root is undefined when $T_i(t-1) = 0$. The value inside the argmax is called the **index** of arm i . Generally speaking, an **index algorithm** chooses the arm in each round that maximizes some value (the index), which

usually only depends on current time-step and the samples from that arm. In the case of UCB, the index is the sum of the empirical mean of rewards experienced so far and the **exploration bonus** (also known as the **confidence width**).

Besides the slightly vague ‘optimism guarantees optimality or learning’ intuition we gave before, it is worth exploring other intuitions for the choice of index. At a very basic level, an algorithm should explore arms more often if they are (a) promising ($\hat{\mu}_i(t-1)$ is large) or (b) not well explored ($T_i(t-1)$ is small). As one can plainly see from the definition, the index above exhibits this behavior. This explanation is not completely satisfying, however, because it does not explain why the form of the functions is just so.

A more refined explanation comes from thinking of what we expect from any reasonable algorithm. Suppose in some round we have played some arm (let’s say arm 1) much more frequently than the others. If we did a good job designing our algorithm we would hope this is the optimal arm. Since we played it so much we can expect that $\hat{\mu}_1(t-1) \approx \mu_1$. To confirm the hypothesis that arm 1 is optimal the algorithm better be highly confident that other arms are indeed worse. This leads very naturally to confidence intervals and the requirement that $T_i(t-1)$ for other arms $i \neq 1$ better be so large that

$$\hat{\mu}_i(t-1) + \sqrt{\frac{2 \log(1/\delta)}{T_i(t-1)}} \leq \mu_1, \quad (7.3)$$

because, at a **confidence level** of $1 - \delta$ this guarantees that μ_i is smaller than μ_1 and if the above inequality did *not* hold, the algorithm would *not* be justified in choosing arm 1 much more often than arm i . Then, planning for (7.3) to hold makes it reasonable to follow the UCB rule as this will eventually guarantee that this inequality holds when arm 1 is indeed optimal and arm i is suboptimal. That this rule is indeed a good one depends on two factors: The first is whether the width of the confidence interval at a given confidence level can be significantly decreased and the second is whether the confidence level is chosen in a reasonable fashion. For now, we will take a leap of faith and assume that the width of confidence intervals for subgaussian bandits cannot be significantly improved from what we use here (we shall see that this holds in later chapters), and concentrate on choosing the confidence level now.

Choosing the confidence level itself turns out to be a delicate problem and we will spend quite a lot of time analyzing various choices in future chapters.



The basic difficulty is that there is a trade-off between choosing δ very small and ensuring optimism with very high probability, and the cost of being excessively optimistic about suboptimal arms. Note that optimism is only really required for the optimal arm because this ensures that once the suboptimal arms have been proven to have means less than optimal, then the optimal arm is all that remains.

Nevertheless, as a first cut, the choice of this parameter can be guided by the following considerations. If the confidence interval fails and the index of an optimal arm drops below its true mean, then it could happen that the algorithm stops playing the optimal arm and suffers linear regret. This suggests we might choose $\delta \approx 1/n$ so that the contribution to the regret of this failure case is relatively small. Unfortunately things are not quite this simple. As we have already alluded to, one of the main difficulties is that the number of samples $T_i(t-1)$ in the index (7.2) is a random variable and so our concentration results cannot be immediately applied. For this reason we will see that (at least naively) δ should be chosen a bit smaller than $1/n$.

THEOREM 7.1 *Consider UCB as shown in Algorithm 2 on a stochastic K-armed 1-subgaussian bandit problem. For any horizon n , if $\delta = 1/n^2$ then*

$$R_n \leq 3 \sum_{i=1}^K \Delta_i + \sum_{i:\Delta_i > 0} \frac{16 \log(n)}{\Delta_i}.$$

Before the proof we need a little more notation. Let $(X_{ti})_{t \in [n], i \in [K]}$ be a collection of independent random variables with the law of X_{ti} equal to P_i . Then define $\hat{\mu}_{is} = \frac{1}{s} \sum_{u=1}^s X_{ui}$ to be the empirical mean based on the first s samples. We make use of the third model in Section 4.4 by assuming that the reward in round t is

$$X_t = X_{A_t T_{A_t}(t)}.$$

Then we define $\hat{\mu}_i(t) = \hat{\mu}_{iT_i(t)}$ to be the empirical mean of the i th arm after round t . The proof of Theorem 7.1 relies on the basic regret decomposition identity,

$$R_n = \sum_{i=1}^K \Delta_i \mathbb{E}[T_i(n)]. \quad (\text{Lemma 4.2})$$

The theorem will follow by showing that $\mathbb{E}[T_i(n)]$ is not too large for suboptimal arms i . The key observation is that after the initial period where the algorithm chooses each action once, action i can only be chosen if its index is higher than that of an optimal arm. This can only happen if at least one of the following is true:

- (a) The index of action i is larger than the true mean of a specific optimal arm.
- (b) The index of a specific optimal arm is smaller than its true mean.

Since with reasonably high probability the index of any arm is an upper bound on its mean, we don't expect the index of the optimal arm to be below its mean. Furthermore, if the suboptimal arm i is played sufficiently often, then its exploration bonus becomes small and simultaneously the empirical estimate of its mean converges to the true value, putting an upper bound on the expected total number of times when its index stays above the mean of the optimal arm. The proof that follows is typical for the analysis of algorithms like UCB and hence we provide quite a bit of detail so that readers can later construct their own proofs.

Proof of Theorem 7.1 Without loss of generality we assume the first arm is optimal so that $\mu_1 = \mu^*$. As noted above,

$$R_n = \sum_{i=1}^K \Delta_i \mathbb{E}[T_i(n)]. \quad (7.4)$$

The theorem will be proven by bounding $\mathbb{E}[T_i(n)]$ for each suboptimal arm i . We make use of a relatively standard idea, which is to decouple the randomness from the behavior of the UCB algorithm. Let G_i be the ‘good’ event defined by

$$G_i = \left\{ \mu_1 < \min_{t \in [n]} \text{UCB}_1(t) \right\} \cap \left\{ \hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i} \log\left(\frac{1}{\delta}\right)} < \mu_1 \right\},$$

where $u_i \in [n]$ is a constant to be chosen later. So G_i is the event when μ_1 is never underestimated by the upper confidence bound of the first arm, while at the same time the upper confidence bound for the mean of arm i after u_i observations are taken from this arm is below the payoff of the optimal arm. We will show two things:

- 1 If G_i occurs, then $T_i(n) \leq u_i$.
- 2 The complement event G_i^c occurs with low probability (governed in some way yet to be discovered by u_i).

Because $T_i(n) \leq n$ no matter what, this will mean that

$$\mathbb{E}[T_i(n)] = \mathbb{E}[\mathbb{I}\{G_i\} T_i(n)] + \mathbb{E}[\mathbb{I}\{G_i^c\} T_i(n)] \leq u_i + \mathbb{P}(G_i^c) n. \quad (7.5)$$

The next step is to complete our promise by showing that $T_i(n) \leq u_i$ on G_i and that $\mathbb{P}(G_i^c)$ is small. Let us first assume that G_i holds and show that $T_i(n) \leq u_i$, which we do by contradiction. Suppose that $T_i(n) > u_i$. Then, arm i was played more than u_i times over the n rounds and so there must exist a round $t \in [n]$ where $T_i(t-1) = u_i$ and $A_t = i$. Using the definition of G_i we have:

$$\begin{aligned} \text{UCB}_i(t-1) &= \hat{\mu}_i(t-1) + \sqrt{\frac{2 \log(1/\delta)}{T_i(t-1)}} && \text{(definition of } \text{UCB}_i(t-1)) \\ &= \hat{\mu}_{iu_i} + \sqrt{\frac{2 \log(1/\delta)}{u_i}} && \text{(since } T_i(t-1) = u_i) \\ &< \mu_1 && \text{(definition of } G_i) \\ &< \text{UCB}_1(t-1), && \text{(definition of } G_i) \end{aligned}$$

which means that $A_t = \text{argmax}_j \text{UCB}_j(t-1) \neq i$ and so a contradiction is obtained. Therefore if G_i occurs, then $T_i(n) \leq u_i$. Let us now turn to upper bounding $\mathbb{P}(G_i^c)$. By its definition,

$$G_i^c = \left\{ \mu_1 \geq \min_{t \in [n]} \text{UCB}_1(t) \right\} \cup \left\{ \hat{\mu}_{iu_i} + \sqrt{\frac{2 \log(1/\delta)}{u_i}} \geq \mu_1 \right\}. \quad (7.6)$$

The first of these sets is decomposed using the definition of $\text{UCB}_1(t)$

$$\begin{aligned} \left\{ \mu_1 \geq \min_{t \in [n]} \text{UCB}_1(t) \right\} &\subset \left\{ \mu_1 \geq \min_{s \in [n]} \hat{\mu}_{1s} + \sqrt{\frac{2 \log(1/\delta)}{s}} \right\} \\ &= \bigcup_{s \in [n]} \left\{ \mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2 \log(1/\delta)}{s}} \right\}. \end{aligned}$$

Then using a union bound and the concentration bound for sums of independent subgaussian random variables in Corollary 5.1 we obtain:

$$\begin{aligned} \mathbb{P} \left(\mu_1 \geq \min_{t \in [n]} \text{UCB}_1(t) \right) &\leq \mathbb{P} \left(\bigcup_{s \in [n]} \left\{ \mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2 \log(1/\delta)}{s}} \right\} \right) \\ &\leq \sum_{s=1}^n \mathbb{P} \left(\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2 \log(1/\delta)}{s}} \right) \leq n\delta. \quad (7.7) \end{aligned}$$

The next step is to bound the probability of the second set in (7.6). Assume that u_i is chosen large enough that

$$\Delta_i - \sqrt{\frac{2 \log(1/\delta)}{u_i}} \geq c\Delta_i \quad (7.8)$$

for some $c \in (0, 1)$ to be chosen later. Then, since $\mu_1 = \mu_i + \Delta_i$ and using Corollary 5.1,

$$\begin{aligned} \mathbb{P} \left(\hat{\mu}_{iu_i} + \sqrt{\frac{2 \log(1/\delta)}{u_i}} \geq \mu_1 \right) &= \mathbb{P} \left(\hat{\mu}_{iu_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2 \log(1/\delta)}{u_i}} \right) \\ &\leq \mathbb{P} (\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i) \leq \exp \left(-\frac{u_i c^2 \Delta_i^2}{2} \right). \end{aligned}$$

Taking this together with (7.7) and (7.6) we have

$$\mathbb{P}(G_i^c) \leq n\delta + \exp \left(-\frac{u_i c^2 \Delta_i^2}{2} \right).$$

When substituted into Eq. (7.5) we obtain

$$\mathbb{E}[T_i(n)] \leq u_i + n \left(n\delta + \exp \left(-\frac{u_i c^2 \Delta_i^2}{2} \right) \right). \quad (7.9)$$

It remains to choose u_i , which must be a positive integer and satisfy (7.8). A natural choice is the smallest integer for which (7.8) holds, which is

$$u_i = \left\lceil \frac{2 \log(1/\delta)}{(1-c)^2 \Delta_i^2} \right\rceil.$$

Then using the assumption that $\delta = 1/n^2$ and this choice of u_i leads via (7.9) to

$$\mathbb{E}[T_i(n)] \leq u_i + 1 + n^{1-2c^2/(1-c)^2} = \left\lceil \frac{2 \log(n^2)}{(1-c)^2 \Delta_i^2} \right\rceil + 1 + n^{1-2c^2/(1-c)^2}. \quad (7.10)$$

All that remains is to choose $c \in (0, 1)$. The second term will contribute a polynomial dependence on n unless $2c^2/(1 - c)^2 \geq 1$. However, if c is chosen too close to 1, then the first term blows up. Somewhat arbitrarily we choose $c = 1/2$, which leads to

$$\mathbb{E}[T_i(n)] \leq 3 + \frac{16 \log(n)}{\Delta_i^2}.$$

The result follows by substituting the above display in Eq. (7.4). \square

As we saw for the ETC strategy, the regret bound in Theorem 7.1 depends on the reciprocal of the gaps, which may be meaningless when even a single suboptimal action has a very small suboptimality gap. As before one can also prove a sublinear regret bound that does not depend on the reciprocal of the gaps.

THEOREM 7.2 *If $\delta = 1/n^2$, then the regret of UCB, as defined in Algorithm 2, on any $\nu \in \mathcal{E}_{\text{SG}}^K(1)$ environment is bounded by*

$$R_n \leq 8\sqrt{nK \log(n)} + 3 \sum_{i=1}^K \Delta_i.$$

Proof Let $\Delta > 0$ be some value to be tuned subsequently and recall from the proof of Theorem 7.1 that for each suboptimal arm i we can bound

$$\mathbb{E}[T_i(n)] \leq 3 + \frac{16 \log(n)}{\Delta_i^2}.$$

Therefore using the basic regret decomposition again (Lemma 4.2), we have

$$\begin{aligned} R_n &= \sum_{i=1}^K \Delta_i \mathbb{E}[T_i(n)] = \sum_{i: \Delta_i < \Delta} \Delta_i \mathbb{E}[T_i(n)] + \sum_{i: \Delta_i \geq \Delta} \Delta_i \mathbb{E}[T_i(n)] \\ &\leq n\Delta + \sum_{i: \Delta_i \geq \Delta} \left(3\Delta_i + \frac{16 \log(n)}{\Delta_i} \right) \leq n\Delta + \frac{16K \log(n)}{\Delta} + 3 \sum_i \Delta_i \\ &\leq 8\sqrt{nK \log(n)} + 3 \sum_{i=1}^K \Delta_i, \end{aligned}$$

where the first inequality follows because $\sum_{i: \Delta_i < \Delta} T_i(n) \leq n$ and the last line by choosing $\Delta = \sqrt{16K \log(n)/n}$. \square

The additive $\sum_i \Delta_i$ term is unavoidable because no reasonable algorithm can avoid playing each arm once (try to work out what would happen if it did not). In any case, this term does not grow with the horizon n and is typically negligible. As it happens, Theorem 7.2 is close to optimal. We will see in Chapter 15 that no algorithm can enjoy regret smaller than $O(\sqrt{nK})$ over all problems in $\mathcal{E}_{\text{SG}}^K(1)$. In Chapter 9 we will also see a more complicated variant of Algorithm 2 that shaves the logarithmic term from the upper bound given above.

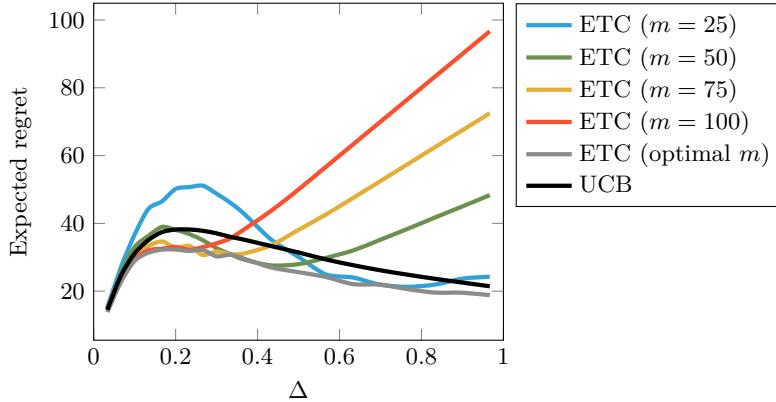


Figure 7.1 Experiment showing universality of UCB relative to fixed instances of explore-then-commit



We promised that UCB would overcome the limitations of ETC by achieving the same guarantees, but without prior knowledge of the suboptimality gaps. The theory supports this claim, but just because two algorithms have similar theoretical guarantees, does not mean they perform the same empirically. The theoretical analysis might be loose for one algorithm (and maybe not the other, or by a different margin). For this reason it is always wise to prove lower bounds (which we do later) and compare the empirical performance, which we do (very briefly) now.

The setup is the same as in Fig. 6.1, which has $n = 1000$ and $K = 2$ and unit variance Gaussian rewards with means 0 and $-\Delta$ respectively. The plot in Fig. 7.1 shows the expected regret of UCB relative to ETC for a variety of choices of commitment time m . The expected regret of ETC with the optimal choice of m (which depends on the knowledge of Δ and that the payoffs are Gaussian, cf. Fig. 6.1) is also shown.

The results demonstrate a common phenomenon. If ETC is tuned with the optimal choice of commitment time for each choice of Δ then it can outperform the parameter-free UCB, though only by a relatively small margin. If, however, the commitment time must be chosen without the knowledge of Δ , for Δ getting large, or for Δ being bounded, n getting large, UCB arbitrarily outperforms ETC. As it happens, a variant of UCB introduced in the next chapter actually outperforms even the optimally tuned ETC.

7.1 Notes

- 1 The choice of $\delta = 1/n^2$ led to an easy analysis, but comes with two disadvantages. First of all, it turns out that a slightly smaller value of δ improves the regret (and empirical performance). Secondly, the dependence on n means the horizon must be known in advance, which is often not reasonable. Both of these issues are resolved in the next chapter where δ is chosen to be smaller and to depend on the current round t rather than n . None-the-less – as promised – Algorithm 2 with $\delta = 1/n^2$ does achieve a regret bound similar to the ETC strategy, but without requiring knowledge of the gaps.
- 2 The assumption that the rewards generated by each arm are independent can be relaxed significantly. All of the results would go through by assuming there exists a mean reward vector $\mu \in \mathbb{R}^K$ such that

$$\mathbb{E}[X_t | X_1, A_1, \dots, A_{t-1}, X_{t-1}, A_t] = \mu_{A_t} \text{ a.s.} \quad (7.11)$$

$$\mathbb{E}[\exp(\lambda(X_t - \mu_{A_t})) | X_1, A_1, \dots, A_{t-1}, X_{t-1}, A_t] \leq \exp(\lambda^2/2) \text{ a.s.} \quad (7.12)$$

Eq. (7.11) is just saying that the conditional mean of the reward in round t only depends on the chosen action. Eq. (7.12) ensures that the tails of X_t are conditionally subgaussian. That everything still goes through is proven using martingale techniques, which we develop in detail in Chapter 20.

- 3 So is the optimism principle universal? Does it always give good algorithms, even in more complicated settings? Unfortunately, the answer is no. The optimism principle leads to reasonable algorithms when (i) any action gives feedback about how much that action is worth, and (ii) no action gives feedback about the value of other actions. When either of these conditions is not met (i.e., in structured bandits, or learning problems with less-than-bandit-feedback such as when one has to choose some action $B \neq A$ to learn about the rewards of some action A), the principle fails! When (i) is violated even sublinear regret may not be guaranteed. When (ii) is violated the regret achieved by optimistic algorithm may be too high (optimistic algorithms may miss some nontrivial optimization opportunities). Thus, unstructured finite-armed stochastic bandits, when both (i) and (ii) hold, are a perfect fit for optimistic algorithms. While the more complex models where the above conditions may not be met may not make much sense at the moment, they will be discussed quite extensively in later chapters.

7.2 Bibliographical remarks

The use of confidence bounds and the idea of optimism first appeared in the work by [Lai and Robbins \[1985\]](#) (for the curious, it is the same Robbins). They analyzed the asymptotics for various parametric bandit problems (see the next chapter for more details on this). The first version of UCB is by [Lai \[1987\]](#). Other

early work is by [Katehakis and Robbins \[1995\]](#), who gave a very straightforward analysis for the Gaussian case and [Agrawal \[1995\]](#), who noticed that all that was needed is an appropriate sequence of upper confidence bounds on the unknown means. In this way, their analysis is significantly more general than what we have done here. These researchers also focussed on the asymptotics, which at the time was the standard approach in the statistics literature. The UCB algorithm was independently discovered by [Kaelbling \[1993\]](#), although with no regret analysis or clear advice on how to tune the confidence parameter. The version of UCB discussed here is most similar to that analyzed by [Auer et al. \[2002a\]](#) under the name UCB1, but that algorithm used t rather than n in the confidence level (see the next chapter). Like us, they prove a finite-time regret bound. However, rather than considering 1-subgaussian environments, [Auer et al. \[2002a\]](#) considers bandits where the payoffs are confined to the $[0, 1]$ interval, which are ensured to be 1/2-subgaussian. See Exercise 7.2 for hints on what must change in this situation. The basic structure of the proof of our Theorem 7.1 is essentially the same as that of Theorem 1 of [Auer et al. \[2002a\]](#). The worst-case bound in Theorem 7.2 appeared in the book by [Bubeck and Cesa-Bianchi \[2012\]](#), which also popularized the subgaussian setup. We did not have time to discuss the situation where the subgaussian constant is unknown. There have been several works exploring this direction. If the variance is unknown, but the noise is bounded, then one can replace the subgaussian concentration bounds with an empirical Bernstein inequality [\[Audibert et al., 2007\]](#). For details see Exercise 7.7. If the noise has heavy tails, then a more serious modification is required as discussed in Exercise 7.8 and the note that follows.

7.3 Exercises

7.1 In this exercise we investigate one of the more annoying challenges when analyzing sequential algorithms. Let X_1, X_2, \dots be a sequence of independent standard Gaussian random variables defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that $T : \Omega \rightarrow \{1, 2, 3, \dots\}$ is another random variable and let $\hat{\mu} = \sum_{t=1}^T X_t / T$ be the empirical mean based on T samples.

- (a) Show that if T is independent from X_t for all t , then

$$\mathbb{P}\left(\hat{\mu} - \mu \geq \sqrt{\frac{2 \log(1/\delta)}{T}}\right) \leq \delta.$$

- (b) We now relax the assumption that T is independent. Let $E_t = \mathbb{I}\{T = t\}$ be the event that $T = t$ and $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ be the σ -algebra generated by the first t samples. Show there exists a T such that for all $t \in \{1, 2, 3, \dots\}$ it

holds that E_t is \mathcal{F}_t -measurable and

$$\mathbb{P} \left(\hat{\mu} - \mu \geq \sqrt{\frac{2 \log(1/\delta)}{T}} \right) = 1 \quad \text{for all } \delta \in (0, 1).$$

(c) Show that

$$\mathbb{P} \left(\hat{\mu} - \mu \geq \sqrt{\frac{2 \log(T(T+1)/\delta)}{T}} \right) \leq \delta. \quad (7.13)$$



For part (b) above you may find it useful to apply the law of the iterated logarithm, which says if X_1, X_2, \dots is a sequence of independent and identically distributed random variables with zero mean and unit variance, then

$$\limsup_{n \rightarrow \infty} \frac{\sum_{t=1}^n X_t}{\sqrt{2n \log \log n}} = 1 \quad \text{almost surely}.$$

This result is especially remarkable because it relies on no assumptions other than zero mean and unit finite variance. A thoughtful reader might wonder if Eq. (7.13) might still be true if $\log(T(T+1))/\delta$ were replaced by $\log(\log(T)/\delta)$. It almost can, but the proof of this fact is more sophisticated. For more details see the paper by [Garivier \[2013\]](#) or Exercise 23.6.

7.2 In this chapter we assumed the payoff distributions were 1-subgaussian (that is, $\nu \in \mathcal{E}_{\text{SG}}^K(1)$). The purpose of this exercise is to relax this assumption.

- (a) First suppose that $\sigma^2 > 0$ is a known constant and that $\nu \in \mathcal{E}_{\text{SG}}^K(\sigma^2)$. Modify the UCB algorithm and state and prove an analogue of Theorems 7.1 and 7.2 for this case.
- (b) Now suppose that $\nu = (\nu_i)_i$ is chosen so that ν_i is σ_i -subgaussian where $(\sigma_i^2)_i$ are known. Modify the UCB algorithm and state and prove an analogue of Theorems 7.1 and 7.2 for this case.
- (c) If you did things correctly, the regret bound in the previous part should not depend on the values of $\{\sigma_i^2 : \Delta_i = 0\}$. Explain why not.

7.3 Recall from Chapter 4 that the pseudo-regret is defined to be the random variable

$$\bar{R}_n = \sum_{t=1}^n \Delta_{A_t}.$$

The UCB policy in Algorithm 2 depends on confidence parameter $\delta \in (0, 1]$ that determines the level of optimism. State and prove a bound on the pseudo-regret of this algorithm that holds with probability $1 - f(n, K)\delta$ where $f(n, K)$ is a function that depends on n and K only. More precisely show that for bandit $\nu \in \mathcal{E}_{\text{SG}}^K(1)$ that

$$\mathbb{P} (\bar{R}_n \geq g(n, \nu, \delta)) \leq f(n, K)\delta,$$

where g and f should be as small as possible (there are trade-offs – try and come up with a natural choice).

7.4 This exercise is about the empirical behavior of UCB.

- (a) Implement Algorithm 2.
- (b) Reproduce Fig. 7.1.
- (c) Explain the shape of the ETC curves. In particular, when $m = 50$ we see a bump, a dip, and then a linear asymptote as Δ grows. Why does the curve look like this?
- (d) Design an experiment to determine the practical effect of the choice of δ . There are many interesting regimes where this is interesting. For example:
 - (1) Suppose you have a Gaussian bandit with two arms and means $\mu_1 = 0$ and $\mu_2 = -\Delta$. Let $n = 1000$ and try to determine the optimal value of δ for UCB as a function of Δ .
 - (2) What happens if you have more arms. For example, $\mu_1 = 0$ and $\mu_i = -\Delta$ for $i > K$. How does the optimal choice of δ change as K increases?
 - (3) Justify your results pseudo-theoretically (that is, provide a theoretically motivated justification for the results, but no proof).

7.5 Fix a 1-subgaussian K -armed bandit environment and a horizon n . Consider the version of UCB that works in phases of exponentially increasing length of $1, 2, 4, \dots$. In each phase, the algorithm uses the action that would have been chosen by UCB at the beginning of the phase (see Algorithm 3 below).

- (a) State and prove a bound on the regret for this version of UCB.
- (b) Compare your result with Theorem 7.1.
- (c) How would the result change if the k th phase had a length of $\lceil \alpha^k \rceil$ with $\alpha > 1$?

```

1: Input  $K$  and  $\delta$ 
2: Choose each arm once
3: for  $\ell = 1, 2, \dots$  do
4:   Compute  $A_\ell = \text{argmax}_i \text{UCB}_i(t - 1, \delta)$ 
5:   Choose arm  $A_\ell$  exactly  $2^\ell$  times
6: end for
```

Algorithm 3: A phased version of UCB

7.6 Let $\alpha > 1$ and consider the version of UCB that first plays each arm once. Thereafter it operates in the same way as UCB, but rather than playing the chosen arm just once, it plays it until the number of plays of that arm is a factor of α larger (see Algorithm 4 below).

- (a) State and prove a bound on the regret for this version of UCB with $\alpha = 2$ (doubling counts).

- (b) Compare with the result of the previous exercise and with Theorem 7.1. What can you conclude?
- (c) Repeat the analysis for $\alpha > 1$. What is the role of α ?
- (d) Implement these algorithms and compare them empirically to UCB(δ).

```

1: Input  $K$  and  $\delta$ 
2: Choose each arm once
3: for  $\ell = 1, 2, \dots$  do
4:   Let  $t_\ell = t$ 
5:   Compute  $A_\ell = \text{argmax}_i \text{UCB}_i(t_\ell - 1, \delta)$ 
6:   Choose arm  $A_\ell$  until round  $t$  such that  $T_i(t) \geq \alpha T_i(t_\ell - 1)$ 
7: end for
```

Algorithm 4: A phased version of UCB

The algorithms of the last two exercises may seem ridiculous. Why would you wait before updating empirical estimates and choosing a new action? There are at least two reasons:

- It can happen that the algorithm does not observe its rewards immediately, but rather they appear asynchronously after some delay. Alternatively many bandits algorithms may be operating simultaneously and the results must be communicated at some cost.
- If the feedback model has a more complicated structure than what we examined so far, then even computing the upper confidence bound just once can be quite expensive. In these circumstances it's comforting to know that the loss of performance by updating the statistics only rarely is not too severe.

7.7 Let X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables with mean μ and variance σ^2 and bounded support so that $X_t \in [0, b]$ almost surely. Let $\hat{\mu} = \sum_{t=1}^n X_t/n$ and $\hat{\sigma}^2 = \sum_{t=1}^n (\hat{\mu} - X_t)^2/n$. The **empirical Bernstein** inequality says that for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(|\hat{\mu} - \mu| \geq \sqrt{\frac{2\hat{\sigma}^2}{n} \log\left(\frac{3}{\delta}\right)} + \frac{3b}{n} \log\left(\frac{3}{\delta}\right)\right) \leq \delta.$$

- (a) Show that $\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^n (X_t - \mu)^2 - (\hat{\mu} - \mu)^2$.
- (b) Show that $\mathbb{V}[(X_t - \mu)^2] \leq b^2 \sigma^2$.
- (c) Use Bernstein's inequality (Exercise 5.16) to show that

$$\mathbb{P}\left(\hat{\sigma}^2 \geq \sigma^2 + \sqrt{\frac{2b^2 \sigma^2}{n} \log\left(\frac{1}{\delta}\right)} + \frac{2b^2}{3n} \log\left(\frac{1}{\delta}\right)\right) \leq \delta.$$

- (d) Suppose that $\nu = (\nu_i)_{i=1}^K$ is a bandit where $\text{Supp}(\nu_i) \subset [0, b]$ and the variance of the i th arm is σ_i^2 . Design a policy that depends on b , but not σ_i^2 such that

$$R_n \leq C \sum_{i:\Delta_i > 0} \left(\Delta_i + \left(b + \frac{\sigma_i^2}{\Delta_i} \right) \log(n) \right),$$

where $C > 0$ is a universal constant.



If you did things correctly, then the policy you derived in Exercise 7.7 should resemble UCB-V by [Audibert et al. \[2007\]](#). The proof of the empirical Bernstein also appears there or in the papers by [Mnih et al. \[2008\]](#) and [Maurer and Pontil \[2009\]](#).

- 7.8** Let $n \in \mathbb{N}^+$ and $(A_i)_{i=1}^k$ be a partition of $[n]$ so that $\cup_{i=1}^k A_i = [n]$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$. Suppose that $\delta \in (0, 1)$ and X_1, X_2, \dots, X_n is a sequence of independent random variables with mean μ and variance σ^2 . The **median-of-means estimator** $\hat{\mu}_M$ of μ the median of $\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_k$ where $\hat{\mu}_i = \sum_{t \in A_i} X_t / |A_i|$ is the mean of the data in the i th block.

- (a) Show that if $k = \left\lfloor \min \left\{ \frac{n}{2}, 8 \log \left(\frac{e^{1/8}}{\delta} \right) \right\} \right\rfloor$ and A_i are chosen as equally sized as possible, then

$$\mathbb{P} \left(\hat{\mu}_M + \sqrt{\frac{192\sigma^2}{n} \log \left(\frac{e^{1/8}}{\delta} \right)} \right) \leq \delta.$$

- (b) Use the median-of-means estimator to design an upper confidence bound algorithm such that for all $\nu \in \mathcal{E}_{\mathbb{V}}^K(\sigma^2)$

$$R_n \leq C \sum_{i:\Delta_i > 0} \left(\Delta_i + \frac{\sigma^2 \log(n)}{\Delta_i} \right),$$

where $C > 0$ is a universal constant.



This exercise shows that unless one cares greatly about constant factors, then the subgaussian assumption can be relaxed to requiring only finite variance. The result is only possible by replacing the standard empirical estimator with something more robust. The median-of-means estimator is only one way to do this. In fact, the empirical estimator can be made robust by truncating the observed rewards and applying the empirical Bernstein concentration inequality. The disadvantage of this approach is that choosing the location of truncation requires prior knowledge about the approximate location of the mean. Another approach is **Catoni's estimator**, which also exhibits excellent asymptotic properties [[Catoni, 2012](#)]. Yet another idea is to minimize the Huber loss [[Sun et al., 2017](#)]. This latter paper is focussing on linear models, but the results still apply in one dimension. The application of these ideas to bandits was first

made by [Bubeck et al. \[2013a\]](#), where the reader will find more interesting results. Most notably, that things can still be made to work even if the variance does not exist. In this case, however, there is a price to be paid in terms of the regret. The median-of-means estimator is due to [Alon et al. \[1996\]](#). In case the variance is also unknown, then it may be estimated by assuming a known bound on the **kurtosis**, which covers many classes of bandits (Gaussian with arbitrary variance, exponential and many more), but not some simple cases (Bernoulli). The policy that results from this procedure has the benefit of being invariant under the transformations of shifting or scaling the losses [[Lattimore, 2017](#)].

8 The Upper Confidence Bound Algorithm: Asymptotic Optimality

In the next few chapters we improve the regret of policies based on the optimism principle by refining the confidence level used by UCB. Our first refinement of Algorithm 2 improves the constants in the previous analysis and resolves the issue of knowing the horizon in advance. As we shall eventually discover in Chapter 16 on lower bounds, the analysis here is sufficiently tight that the dominant logarithmic terms in the regret bound are preceded by the optimal constants.

- 1: **Input** K
- 2: Choose each arm once
- 3: Subsequently choose

$$A_t = \operatorname{argmax}_i \left(\hat{\mu}_i(t-1) + \sqrt{\frac{2 \log f(t)}{T_i(t-1)}} \right)$$

$$\text{where } f(t) = 1 + t \log^2(t)$$

Algorithm 5: Asymptotically optimal UCB

The regret bound for Algorithm 5 is more complicated than what we presented for Algorithm 2 (see Theorem 7.1). The important thing is that the dominant terms have the same order and slightly smaller constants.

THEOREM 8.1 *The regret of Algorithm 5 satisfies*

$$R_n \leq \sum_{i: \Delta_i > 0} \inf_{\varepsilon \in (0, \Delta_i)} \Delta_i \left(1 + \frac{5}{\varepsilon^2} + \frac{2}{(\Delta_i - \varepsilon)^2} (\log f(n) + \sqrt{\pi \log f(n)} + 3) \right). \quad (8.1)$$

Furthermore,

$$\limsup_{n \rightarrow \infty} \frac{R_n}{\log(n)} \leq \sum_{i: \Delta_i > 0} \frac{2}{\Delta_i}. \quad (8.2)$$

Before the proof, let us present a simpler version of the above bound, avoiding all these epsilons and infimums that make for a confusing theorem statement. By

choosing $\varepsilon = \Delta_i/2$ we see that the regret of Algorithm 5 is bounded by

$$R_n \leq \sum_{i:\Delta_i > 0} \left(\Delta_i + \frac{1}{\Delta_i} \left(8 \log f(n) + 8\sqrt{\pi \log f(n)} + 44 \right) \right). \quad (8.3)$$

Even more concretely, there exists some universal constant $C > 0$ such that

$$R_n \leq C \sum_{i:\Delta_i > 0} \left(\Delta_i + \frac{\log(n)}{\Delta_i} \right),$$

which by the same argument as in the proof of Theorem 7.2 leads a worst-case bound of $R_n \leq C \sum_{i=1}^K \Delta_i + 2\sqrt{CnK \log(n)}$.



Taking the limit of the ratio of the bound in (8.3) and $\log(n)$ does not result in the same constant as in the theorem, which is the main justification for introducing the epsilons in the first place. We shall see in Chapter 15 that the asymptotic bound on the regret given in (8.2), which is derived from (8.1) by choosing $\varepsilon = \log^{-1/4}(n)$, is unimprovable in a strong sense.

We start with a useful lemma that helps us bound the number of times the index of a suboptimal arm will be larger than some thresh-hold above its mean.

LEMMA 8.1 *Let X_1, X_2, \dots be a sequence of independent 1-subgaussian random variables, $\hat{\mu}_t = \frac{1}{t} \sum_{s=1}^t X_s$, $\varepsilon > 0$ and*

$$\kappa = \sum_{t=1}^n \mathbb{I} \left\{ \hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \varepsilon \right\}, \quad \kappa' = u + \sum_{t=\lceil u \rceil}^n \mathbb{I} \left\{ \hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \varepsilon \right\},$$

where $u = 2a\varepsilon^{-2}$. Then it holds $\mathbb{E}[\kappa] \leq \mathbb{E}[\kappa'] \leq 1 + \frac{2}{\varepsilon^2}(a + \sqrt{\pi a} + 1)$.

The intuition for this result is as follows. Since the X_i are 1-subgaussian and independent we have $\mathbb{E}[\hat{\mu}_t] = 0$, so we cannot expect $\hat{\mu}_t + \sqrt{2a/t}$ to be smaller than ε until t is at least $2a/\varepsilon^2$. The lemma confirms that this is indeed of the right order as an estimate for $\mathbb{E}[\kappa]$.

Proof By Corollary 5.1 we have

$$\begin{aligned} \mathbb{E}[\kappa] &\leq \mathbb{E}[\kappa'] = u + \sum_{t=\lceil u \rceil}^n \mathbb{P} \left(\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \varepsilon \right) \leq u + \sum_{t=\lceil u \rceil}^n \exp \left(-\frac{t \left(\varepsilon - \sqrt{\frac{2a}{t}} \right)^2}{2} \right) \\ &\leq 1 + u + \int_u^\infty \exp \left(-\frac{t \left(\varepsilon - \sqrt{\frac{2a}{t}} \right)^2}{2} \right) dt = 1 + \frac{2}{\varepsilon^2}(a + \sqrt{\pi a} + 1) \end{aligned}$$

as required. \square

Proof of Theorem 8.1 As usual, we start with the basic regret decomposition.

$$R_n = \sum_{i:\Delta_i > 0} \Delta_i \mathbb{E}[T_i(n)].$$

The rest of the proof revolves around bounding $\mathbb{E}[T_i(n)]$. Let i be the index of some sub-optimal arm (so that $\Delta_i > 0$). The main idea is to decompose $T_i(n)$ into two terms. The first measures the number of times the index of the optimal arm is less than $\mu_1 - \varepsilon$. The second term measures the number of times that $A_t = i$ and its index is larger than $\mu_1 - \varepsilon$.

$$\begin{aligned} T_i(n) &= \sum_{t=1}^n \mathbb{I}\{A_t = i\} \\ &\leq \sum_{t=1}^n \mathbb{I}\left\{\hat{\mu}_1(t-1) + \sqrt{\frac{2 \log f(t)}{T_1(t-1)}} \leq \mu_1 - \varepsilon\right\} \\ &\quad + \sum_{t=1}^n \mathbb{I}\left\{\hat{\mu}_i(t-1) + \sqrt{\frac{2 \log f(t)}{T_i(t-1)}} \geq \mu_1 - \varepsilon \text{ and } A_t = i\right\}. \end{aligned} \quad (8.4)$$

The proof of the first part of the theorem is completed by bounding the expectation of each of these two sums. Starting with the first, we again use Corollary 5.1:

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^n \mathbb{I}\left\{\hat{\mu}_1(t-1) + \sqrt{\frac{2 \log f(t)}{T_1(t-1)}} \leq \mu_1 - \varepsilon\right\}\right] \\ &= \sum_{t=1}^n \mathbb{P}\left(\hat{\mu}_1(t-1) + \sqrt{\frac{2 \log f(t)}{T_1(t-1)}} \leq \mu_1 - \varepsilon\right) \\ &\leq \sum_{t=1}^n \sum_{s=1}^n \mathbb{P}\left(\hat{\mu}_{1,s} + \sqrt{\frac{2 \log f(t)}{s}} \leq \mu_1 - \varepsilon\right) \\ &\leq \sum_{t=1}^n \sum_{s=1}^n \exp\left(-\frac{s\left(\sqrt{\frac{2 \log f(t)}{s}} + \varepsilon\right)^2}{2}\right) \\ &\leq \sum_{t=1}^n \frac{1}{f(t)} \sum_{s=1}^n \exp\left(-\frac{s\varepsilon^2}{2}\right) \leq \frac{5}{\varepsilon^2}. \end{aligned}$$

The first inequality follows from the union bound over all possible values of $T_1(t-1)$. The last inequality is an algebraic exercise (cf. Exercise 8.1). The function $f(t)$ was chosen precisely so this bound would hold. If $f(t) = t$ instead, then the sum would diverge. Since $f(n)$ appears in the numerator below we would like f to be large enough that its reciprocal is summable and otherwise as small

as possible. For the second term in (8.4) we use Lemma 8.1 to get

$$\begin{aligned}
& \mathbb{E} \left[\sum_{t=1}^n \mathbb{I} \left\{ \hat{\mu}_i(t-1) + \sqrt{\frac{2 \log f(t)}{T_i(t-1)}} \geq \mu_1 - \varepsilon \text{ and } A_t = i \right\} \right] \\
& \leq \mathbb{E} \left[\sum_{t=1}^n \mathbb{I} \left\{ \hat{\mu}_i(t-1) + \sqrt{\frac{2 \log f(n)}{T_i(t-1)}} \geq \mu_1 - \varepsilon \text{ and } A_t = i \right\} \right] \\
& \leq \mathbb{E} \left[\sum_{s=1}^n \mathbb{I} \left\{ \hat{\mu}_{i,s} + \sqrt{\frac{2 \log f(n)}{s}} \geq \mu_1 - \varepsilon \right\} \right] \\
& = \mathbb{E} \left[\sum_{s=1}^n \mathbb{I} \left\{ \hat{\mu}_{i,s} - \mu_i + \sqrt{\frac{2 \log f(n)}{s}} \geq \Delta_i - \varepsilon \right\} \right] \\
& \leq 1 + \frac{2}{(\Delta_i - \varepsilon)^2} \left(\log f(n) + \sqrt{\pi \log f(n)} + 1 \right).
\end{aligned}$$

The first part of the theorem follows by substituting the results of the previous two displays into (8.4). The second part follows by choosing $\varepsilon = \log^{-1/4}(n)$ and taking the limit as n tends to infinity. \square

8.1 Notes

- 1 The improvement to the constants comes from making the confidence interval slightly smaller, which is made possible by a more careful analysis. The main trick is the observation that we do not need to show that $\hat{\mu}_{1,s} \geq \mu_1$ for all s with high probability, but instead that $\hat{\mu}_{1,s} \geq \mu_1 - \varepsilon$ for small ε . This idea buys quite a lot and we will see it repeatedly in subsequent chapters.
- 2 The choice of $f(t) = 1 + t \log^2(t)$ looks quite odd. As we pointed out in the proof, things would not have gone through had we chosen $f(t) = t$. With a slightly messier calculation we could have chosen $f(t) = t \log^\alpha(t)$ for any $\alpha > 0$. If the rewards are actually Gaussian, then a more careful concentration analysis allows one to choose $f(t) = t$ or even slightly smaller [Katehakis and Robbins, 1995, Lattimore, 2016a, Garivier et al., 2016b].

8.2 Bibliographic remarks

Lai and Robbins [1985] designed policies for which Eq. (8.2) held and proved lower bounds showing that no ‘reasonable’ policy can improve on this bound for any problem, where ‘reasonable’ means that they suffer subpolynomial regret on all problems. We will discuss these issues in great detail in Part IV where we address lower bounds. The policy proposed by Lai and Robbins [1985] was based on upper confidence bounds, but was not a variant of UCB. The asymptotics for variants of the policy presented here were given first by Katehakis and Robbins [1995] and

Agrawal [1995]. Neither of these articles gave finite-time bounds like what was presented here. When the reward distributions lie in an exponential family, then asymptotic and finite-time bounds with the same flavor to what is presented here are given by Cappé et al. [2013]. There are now a huge variety of asymptotically optimal policies in a wide range of settings. Burnetas and Katehakis [1996] study the general case and give conditions for a version of UCB to be asymptotically optimal. Honda and Takemura [2010, 2011] analyze an algorithm called DMED to derive asymptotic optimality for noise models where the support is bounded or semi-bounded. Kaufmann et al. [2012b] prove asymptotic optimality for Thompson sampling (see Chapter 35) when the rewards are Bernoulli, which is generalized to single parameter exponential families by Korda et al. [2013]. Kaufmann [2018] proves asymptotic optimality for the Bayes UCB class of algorithms for single parameter exponential families. Ménard and Garivier [2017] prove asymptotic optimality and minimax optimality for exponential families (more discussion in Chapter 9).

8.3 Exercises

8.1 [Do the algebra needed at the end of the proof of Theorem 8.1] Show that

$$\sum_{t=1}^n \frac{1}{f(t)} \sum_{s=1}^n \exp\left(-\frac{s\varepsilon^2}{2}\right) \leq \frac{5}{\varepsilon^2},$$

where $f(t) = 1 + t \log^2(t)$.



First bound $F = \sum_{s=1}^n \exp(-s\varepsilon^2/2)$ using a geometric series. Then show that $\exp(-a)/(1 - \exp(-a)) \leq 1/a$ holds for any $a > 0$ and conclude that $F \leq \frac{2}{\varepsilon^2}$. Finish by bounding $\sum_{t=1}^n 1/f(t)$ using the fact that $1/f(t) \leq 1/(t \log(t)^2)$ and bounding a sum by an integral.

8.2 [One-armed bandits and UCB] Consider the one-armed bandit problem from Exercise 4.9. Notice that this one-armed bandit problem can be formulated as a regular bandit with two actions in which the first action corresponds to playing the machine and the second to not playing it. The noise in this case is 1-subgaussian, which means that the theoretical guarantees of UCB are applicable. For $p = 1$, evaluate

$$\limsup_{n \rightarrow \infty} \frac{R_p^{\text{UCB}}(n)}{\log(n)}.$$

8.3 [Continuation of Exercise 8.2] The difference between the one and two-armed bandit is that for one-armed bandits the mean of the second arm is known. This additional information is not exploited by UCB. However, we can incorporate this

additional information into the definition of UCB as follows: Let $f(t) = 1 + t \log^2(t)$ and define a policy by

$$A_t = \begin{cases} 1, & \text{if } \hat{\mu}_1(t-1) + \sqrt{\frac{2 \log f(t)}{T_1(t-1)}} \geq 0; \\ 2, & \text{otherwise.} \end{cases} \quad (8.5)$$

Prove that the modified UCB algorithm satisfies:

$$\limsup_{n \rightarrow \infty} \frac{R_p^{\text{MODIFIED-UCB}}(n)}{\log(n)} \leq \begin{cases} 0, & \text{if } p \geq 1/2; \\ \frac{2}{1-2p}, & \text{if } p < 1/2. \end{cases}$$

(Hint: Follow the analysis that we gave for UCB, but carefully adapt the proof by using the fact that the index of the second arm is always 0. This will leave you with a finite-time regret guarantee for the modified UCB from which the identity above can be derived.)

8.4 [Continuation of Exercise 8.3] The purpose of this question is to compare UCB and the modified version in (8.5).

- (a) Implement a simulator for the one-armed bandit problem and two algorithms. UCB and the modified version analysed in Exercise 8.3.
- (b) Use your simulator to estimate the expected regret of each algorithm for a horizon of $n = 1000$ and $p \in \{0, 1/20, 2/20, \dots, 19/20, 1\}$.
- (c) Plot your results with p on the x -axis and the estimated expected regret on the y -axis. Don't forget to label the axis and include error bars and a legend.
- (d) Explain the results. Why do the curves look the way they do?
- (e) In your plot, for what values of p does the worst-case expected regret for each algorithm occur? What is the worst-case expected regret for each algorithm?

8.5 Let $\sigma^2 \in [0, \infty)^K$ be known and suppose that the reward is $X_t \sim \mathcal{N}(\mu_{A_t}, \sigma_{A_t}^2)$. Design an algorithm (that depends on σ^2) for which the asymptotic regret is

$$\limsup_{n \rightarrow \infty} \frac{R_n}{\log(n)} = \sum_{i: \Delta_i > 0} \frac{2\sigma_i^2}{\Delta_i}.$$

9 The Upper Confidence Bound Algorithm: Minimax Optimality (\dagger)

The variants of UCB analyzed in the last two chapters have a distribution free regret bound of $R_n = O(\sqrt{Kn \log(n)})$. The factor of $\sqrt{\log(n)}$ can be removed by modifying the confidence level of the algorithm. The directly named Minimax Optimal Strategy in the Stochastic case algorithm (MOSS) was the first to make this modification and is presented below. MOSS again depends on prior knowledge of the horizon, a requirement that may be relaxed as we explain in the notes.

 The term **minimax** is used because, except for constant factors, the distribution free bound proven in this chapter cannot be improved upon by any algorithm. The lower bounds are deferred to Part IV.

- 1: **Input** n and K
- 2: Choose each arm once
- 3: Subsequently choose

$$A_t = \operatorname{argmax}_i \hat{\mu}_i(t-1) + \sqrt{\frac{4}{T_i(t-1)} \log^+ \left(\frac{n}{KT_i(t-1)} \right)},$$

where $\log^+(x) = \log \max \{1, x\}$.

Algorithm 6: MOSS

THEOREM 9.1 *The regret of Algorithm 6 is bounded by $R_n \leq 34\sqrt{Kn} + \sum_{i=1}^K \Delta_i$.*

Before the proof we need a strengthened version of Corollary 5.1.

THEOREM 9.2 *Let X_1, X_2, \dots, X_n be a sequence of independent 1-subgaussian random variables and $S_t = \sum_{s=1}^t X_s$. Then,*

$$\mathbb{P}(\exists t \leq n : S_t \geq \varepsilon) \leq \exp \left(-\frac{\varepsilon^2}{2n} \right). \quad (9.1)$$

The bound in Eq. (9.1) is the same as the bound on $\mathbb{P}(S_n \geq \varepsilon)$ that appears in a simple reformulation of Corollary 5.1, so this new result is strictly stronger.

Proof From the definition of subgaussian random variables and Lemma 5.2,

$$\mathbb{E}[\exp(\lambda S_n)] \leq \exp\left(\frac{n\sigma^2\lambda^2}{2}\right).$$

Then choosing $\lambda = \varepsilon/(n\sigma^2)$ leads to

$$\begin{aligned} \mathbb{P}(\text{exists } t \leq n : S_t \geq \varepsilon) &= \mathbb{P}\left(\max_{t \leq n} \exp(\lambda S_t) \geq \exp(\lambda \varepsilon)\right) \\ &\leq \frac{\mathbb{E}[\exp(\lambda S_n)]}{\exp(\lambda \varepsilon)} \leq \exp\left(\frac{n\sigma^2\lambda^2}{2} - \lambda \varepsilon\right) = \exp\left(-\frac{\varepsilon^2}{2n\sigma^2}\right). \end{aligned}$$

The novel step is the first inequality, which follows from the maximal inequality (Theorem 3.5) and the fact that $\exp(\lambda S_t)$ is a supermartingale with respect to the filtration generated by X_1, X_2, \dots, X_n (Exercise 9.1). \square

Before the proof of Theorem 9.1 we need one more lemma to bound the probability that the index of the optimal arm ever drops too far below the actual mean of the optimal arm. The proof of this lemma relies on a tool called the **peeling device**, which is an important technique in probability theory and has many applications beyond bandits. For example, it can be used to prove the law of the iterated logarithm.

LEMMA 9.1 *Let $\delta \in (0, 1)$ and X_1, X_2, \dots be independent and 1-subgaussian and $\hat{\mu}_t = \frac{1}{t} \sum_{s=1}^t X_s$. Then for any $\Delta > 0$,*

$$\mathbb{P}\left(\text{exists } s \geq 1 : \hat{\mu}_s + \sqrt{\frac{4}{s} \log^+\left(\frac{1}{s\delta}\right)} + \Delta \leq 0\right) \leq \frac{16\delta}{\Delta^2}.$$

Proof Let $S_t = t\hat{\mu}_t$. Then

$$\begin{aligned} &\mathbb{P}\left(\text{exists } s \geq 1 : \hat{\mu}_s + \sqrt{\frac{4}{s} \log^+\left(\frac{1}{s\delta}\right)} + \Delta \leq 0\right) \\ &= \mathbb{P}\left(\text{exists } s \geq 1 : S_s + \sqrt{4s \log^+\left(\frac{1}{s\delta}\right)} + s\Delta \leq 0\right) \\ &\leq \sum_{k=0}^{\infty} \mathbb{P}\left(\text{exists } s \in [2^k, 2^{k+1}] : S_s + \sqrt{4s \log^+\left(\frac{1}{s\delta}\right)} + s\Delta \leq 0\right) \\ &\leq \sum_{k=0}^{\infty} \mathbb{P}\left(\text{exists } s \leq 2^{k+1} : S_s + \sqrt{4 \cdot 2^k \log^+\left(\frac{1}{2^{k+1}\delta}\right)} + 2^k \Delta \leq 0\right) \\ &\leq \sum_{k=0}^{\infty} \exp\left(-\frac{\left(\sqrt{2^{k+2} \log^+\left(\frac{1}{2^{k+1}\delta}\right)} + 2^k \Delta\right)^2}{2^{k+2}}\right). \end{aligned}$$

In the first inequality we used the union bound, but rather than applying it on every time step as we did in the proof of Theorem 8.1, we apply it on a geometric

grid. The second step is straightforward, but important because it sets up to apply Theorem 9.2. The rest is purely algebraic.

$$\begin{aligned} \sum_{k=0}^{\infty} \exp \left(-\frac{\left(\sqrt{2^{k+2} \log^+ \left(\frac{1}{2^{k+1} \delta} \right)} + 2^k \Delta \right)^2}{2^{k+2}} \right) &\leq \delta \sum_{k=0}^{\infty} 2^{k+1} \exp(-\Delta^2 2^{k-2}) \\ &\leq \frac{8\delta}{\Delta^2} + \int_0^{\infty} 2^{s+1} \exp(-\Delta^2 2^{s-2}) ds \leq \frac{16\delta}{\Delta^2}, \end{aligned}$$

where the first inequality follows since $(a+b)^2 \geq a^2 + b^2$ for $a, b \geq 0$ and the second last step follows by noting that the integrand is unimodal and has a maximum value of $8\delta/\Delta^2$. For such functions f one may bound $\sum_{k=a}^b f(k) \leq \max_{s \in [a,b]} f(s) + \int_a^b f(s) ds$. \square

Proof of Theorem 9.1 As usual, we assume without loss of generality that the first arm is optimal, so $\mu_1 = \mu^*$. Define a random variable Δ that measures how far below the index of the optimal arm drops below its true mean.

$$\Delta = \left(\mu_1 - \min_{s \leq n} \left(\hat{\mu}_{1s} + \sqrt{\frac{4}{s} \log^+ \left(\frac{n}{Ks} \right)} \right) \right)^+.$$

Using the basic regret decomposition (Lemma 4.2) and splitting the actions based on whether or not their suboptimality gap is smaller or larger than 2Δ leads to

$$\begin{aligned} R_\nu(n) &= \sum_{i: \Delta_i > 0} \Delta_i \mathbb{E}[T_i(n)] \leq 2n\Delta + \sum_{i: \Delta_i > 2\Delta} \Delta_i \mathbb{E}[T_i(n)] \\ &\leq 2n\Delta + 8\sqrt{Kn} + \sum_{i: \Delta_i > \max\{2\Delta, 8\sqrt{K/n}\}} \Delta_i \mathbb{E}[T_i(n)]. \end{aligned}$$

The first term is easily bounded using Proposition 2.3 and Lemma 9.1.

$$\mathbb{E}[2n\Delta] = 2n\mathbb{E}[\Delta] = 2n \int_0^\infty \mathbb{P}(\Delta \geq x) dx \leq 2n \int_0^\infty \min \left\{ 1, \frac{16K}{nx^2} \right\} dx = 16\sqrt{Kn}.$$

For suboptimal arm i define

$$\kappa_i = \sum_{s=1}^n \mathbb{I} \left\{ \hat{\mu}_{is} + \sqrt{\frac{4}{s} \log^+ \left(\frac{n}{Ks} \right)} \geq \mu_i + \Delta_i/2 \right\}.$$

The reason for choosing κ_i in this way is that for arms i with $\Delta_i > 2\Delta$ it holds that the index of the optimal arm is always larger than $\mu_i + \Delta_i/2$ so κ_i is an upper bound on the number of times arm i is played, $T_i(n)$. If $\Delta_i \geq 8(K/n)^{1/2}$,

then the expectation of $\Delta_i \kappa_i$ is bounded using Lemma 8.1 by

$$\begin{aligned}\Delta_i \mathbb{E}[\kappa_i] &\leq \frac{1}{\Delta_i} + \Delta_i \mathbb{E} \left[\sum_{s=1}^n \mathbb{I} \left\{ \hat{\mu}_{is} + \sqrt{\frac{4}{s} \log^+ \left(\frac{n\Delta_i^2}{K} \right)} \geq \mu_i + \Delta_i/2 \right\} \right] \\ &\leq \Delta_i + \frac{8}{\Delta_i} \left(2 \log^+ \left(\frac{n\Delta_i^2}{K} \right) + \sqrt{2\pi \log^+ \left(\frac{n\Delta_i^2}{K} \right)} + 2 \right) \\ &\leq \Delta_i + \sqrt{\frac{n}{K}} \left(2 \log 8 + \sqrt{2\pi \log 8} + 2 \right) \leq \Delta_i + 10\sqrt{\frac{n}{K}},\end{aligned}$$

where the first inequality follows by replacing the s in the logarithm with $1/\Delta_i^2$ and adding the $\Delta_i \times 1/\Delta_i^2$ correction term to compensate for the first Δ_i^{-2} rounds where this doesn't actually hold. Then we use Lemma 8.1 and the monotonicity of $x \rightarrow 1/x \log^+(ax^2)$ for $p \in [0, 1]$ and $ax^2 \geq e^2$. The last inequality follows by naively bounding $2 \log 8 + \sqrt{2\pi \log 8} + 2 \leq 10$. Then

$$\begin{aligned}\sum_{i: \Delta_i > \max\{2\Delta, 8\sqrt{K/n}\}} \Delta_i \mathbb{E}[T_i(n)] &\leq \sum_{i: \Delta_i > \max\{2\Delta, 8\sqrt{K/n}\}} \Delta_i \mathbb{E}[\kappa_i] \\ &\leq \sum_{i: \Delta_i > \max\{2\Delta, 8\sqrt{K/n}\}} \left(\Delta_i + 10\sqrt{\frac{n}{K}} \right) \leq 10\sqrt{nK} + \sum_{i=1}^K \Delta_i.\end{aligned}$$

Combining all the results we have $R_n \leq 34\sqrt{Kn} + \sum_{i=1}^K \Delta_i$. \square

9.1 Notes

- 1 One may also prove an asymptotic upper bound on the regret of MOSS that is rather close to optimal. Specifically, one can show that

$$\limsup_{n \rightarrow \infty} \frac{R_n}{\log(n)} \leq \sum_{i: \Delta_i > 0} \frac{4}{\Delta_i}.$$

By modifying the algorithm slightly it is even possible to replace the 4 with a 2 and so recover the optimal asymptotic regret. The trick is to increase g slightly and replace the 4 in the exploration bonus by 2. The major task is then to re-prove Lemma 9.1, which is done by replacing the intervals $[2^k, 2^{k+1}]$ with smaller intervals $[\xi^k, \xi^{k+1}]$ where ξ is tuned subsequently to be fractionally larger than 1. This procedure is explained in detail by Garivier [2013]. When the reward distributions are actually Gaussian there is a more elegant technique that avoids peeling altogether (Exercise 9.4).

- 2 Although it is not obvious from the bounds proven in this chapter, all versions of MOSS can be arbitrarily worse than UCB in some regimes. This unpleasantness is hidden by both the minimax and asymptotic optimality criteria, which highlights the importance of fully finite-time upper and lower bounds. The counter-example witnessing the failure is quite simple. Let the rewards for all

arms be Gaussian with unit variance and $n = K^3$, $\mu_1 = 0$, $\mu_2 = -\sqrt{K/n}$ and $\mu_i = -1$ for all $i > 2$. From Theorem 8.1 we have that

$$R_n^{\text{UCB}} = O(K \log K),$$

while it turns out that MOSS has a regret of

$$R_n^{\text{MOSS}} = \Omega(\sqrt{Kn}) = \Omega(K^2).$$

A rigorous proof of this claim is quite delicate, but we encourage readers to try to understand why it holds intuitively.

- 3 The easy way to deal with this problem is to replace the index used by MOSS with a less aggressive confidence level.

$$\hat{\mu}_i(t-1) + \sqrt{\frac{4}{T_i(t-1)} \log^+ \left(\frac{n}{T_i(t-1)} \right)}. \quad (9.2)$$

The resulting algorithm is never worse than UCB and you will show in Exercise 9.3 that it has a distribution free regret of $O(\sqrt{nK \log(K)})$. An algorithm that does almost the same thing in disguise is called Improved UCB, which operates in phases and eliminates arms for which the upper confidence bound drops below a lower confidence bound for some arm [Auer and Ortner, 2010]. In practice this algorithm does not perform very well and it is not asymptotically optimal, but the analysis highlights the role of the confidence level in the regret.

- 4 Overcoming the weakness of MOSS without sacrificing minimax optimality is possible by using an adaptive confidence level that tunes the amount of optimism to match the instance. One of the authors has proposed two ways to do this using the following indices.

$$\begin{aligned} \hat{\mu}_i(t-1) + \sqrt{\frac{2(1+\varepsilon)}{T_i(t-1)} \log \left(\frac{n}{t} \right)}. \\ \hat{\mu}_i(t-1) + \sqrt{\frac{2}{T_i(t-1)} \log \left(\frac{n}{\sum_{j=1}^K \min\{T_i(t-1), \sqrt{T_i(t-1)T_j(t-1)}\}} \right)}. \end{aligned} \quad (9.3)$$

The first of these algorithms is called the Optimally Confidence UCB [Lattimore, 2015b] while the second is UCB† Lattimore [2018]. Both algorithms are minimax optimal up to constant factors and never worse than UCB. The latter is also asymptotically optimal. If the horizon is unknown, then UCB† can be modified by replacing n with t . It remains a challenge to provide a straightforward analysis for these algorithms.

- 5 There is a hidden cost of pushing too hard to reduce the expected regret, which is that the variance of the regret can grow significantly. We will analyze this trade-off formally in a future chapter, but explain now the intuition. Consider the two-armed case with suboptimality gap Δ and Gaussian noise. Then the

regret of a carefully tuned algorithm is approximately

$$R_n = O \left(n\Delta\delta + \frac{1}{\Delta} \log \left(\frac{1}{\delta} \right) \right),$$

where δ is a parameter of the policy that determines the likelihood that the optimal arm is misidentified. The choice of δ that minimizes the expected regret depends on Δ and is approximately $1/(n\Delta^2)$. With this choice the regret is

$$R_n = O \left(\frac{1}{\Delta} (1 + \log(n\Delta^2)) \right).$$

Of course Δ is not known in advance, but it can be estimated online so that the above bound is actually realizable by an adaptive policy that does not know Δ in advance (Exercise 9.3). The problem is that with the above choice the second moment of the regret will be at least $\delta(n\Delta)^2 = n$, which is uncomfortably large. On the other hand, choosing $\delta = (n\Delta)^{-2}$ leads to a marginally larger regret of

$$R_n = O \left(\frac{1}{\Delta} \left(\frac{1}{n} + \log(n^2\Delta^2) \right) \right).$$

The second moment for this choice, however, is $O(\log^2(n))$. A discussion of these issues, including empirical results, may also be found in the article by [Audibert et al. \[2007\]](#).

9.2 Bibliographic remarks

The MOSS algorithm is due to [Audibert and Bubeck \[2009\]](#), while an anytime modification is by [Degenne and Perchet \[2016\]](#). The proof that MOSS is asymptotically optimal may be found in the article by [Ménard and Garivier \[2017\]](#). Optimally Confidence UCB and its friends are by one of the authors [Lattimore \[2015b, 2016b, 2018\]](#). The idea to modify the confidence level has been seen in several places, with the earliest by [Lai \[1987\]](#) and more recently by [Honda and Takemura \[2010\]](#). [Kaufmann \[2018\]](#) also used a confidence level like in Eq. (9.2) to derive an algorithm based on Bayesian upper confidence bounds.

9.3 Exercises

9.1 Let X_1, X_2, \dots, X_n be adapted to filtration $\mathbb{F} = (\mathcal{F}_t)_t$ with $\mathbb{E}[X_t | \mathcal{F}_{t-1}] = 0$ almost surely. Prove that $M_t = \exp(\lambda \sum_{s=1}^t X_s)$ is a \mathbb{F} -supermartingale for any $\lambda \in \mathbb{R}$.

9.2 Let $\Delta_{\min} = \min_{i:\Delta_i>0} \Delta_i$. Show there exists a universal constant $C > 0$ such

that the regret of MOSS is bounded by

$$R_n \leq \frac{CK}{\Delta_{\min}} \log^+ \left(\frac{n\Delta_{\min}^2}{K} \right) + \sum_{i=1}^K \Delta_i.$$

9.3 Suppose we modify the index used by MOSS to be

$$\hat{\mu}_i(t-1) + \sqrt{\frac{4}{T_i(t-1)} \log^+ \left(\frac{n}{T_i(t-1)} \right)}.$$

(a) Show that for all 1-subgaussian bandits this new policy suffers regret at most

$$R_n \leq C \left(\sum_{i:\Delta_i > 0} \Delta_i + \frac{1}{\Delta_i} \log^+(n\Delta_i^2) \right),$$

where $C > 0$ is a universal constant.

(b) Under the same conditions as the previous part show there exists a universal constant $C > 0$ such that

$$R_n \leq C \sqrt{Kn \log(K)} + \sum_{i=1}^K \Delta_i.$$

(c) Repeat parts (a) and (b) using the index

$$\hat{\mu}_i(t-1) + \sqrt{\frac{4}{T_i(t-1)} \log^+ \left(\frac{t}{T_i(t-1)} \right)}.$$

9.4 Let $g(t) = at + b$ with $b > 0$ and

$$u(x, t) = \frac{1}{\sqrt{2\pi t}} \exp \left(-\frac{x^2}{2t} \right) - \frac{1}{\sqrt{2\pi t}} \exp \left(-2ab - \frac{(x-2b)^2}{2t} \right)$$

(a) Show that $u(x, t) > 0$ for $x \in (-\infty, g(t))$ and $u(x, t) = 0$ for $x = g(t)$.

(b) Show that $u(x, t)$ satisfies the heat equation:

$$\frac{\partial}{\partial t} u(x, t) = \frac{1}{2} \frac{\partial^2}{\partial x^2} u(x, t).$$

(c) Let B_t be a standard Brownian motion, which for any fixed t has density with respect to the Lebesgue measure.

$$p_t(x) = \frac{1}{\sqrt{2\pi t}} \exp \left(-\frac{x^2}{2t} \right).$$

Define $\tau_g = \min\{t : B_t = g(t)\}$ be the first time the Brownian motion hits the boundary. Put on your physicists hat (or work hard) to argue that

$$\mathbb{P}(\tau_g \geq t) = \int_{-\infty}^{g(t)} u(x, t) dx.$$

- (d) Let $q_g(t)$ be the density of time τ with respect to the Lebesgue measure so that $\mathbb{P}(\tau_g \leq t) = \int_0^t q_g(t)dt$. Show that

$$q_g(t) = \frac{g(0)}{\sqrt{2\pi t^3}} \exp\left(-\frac{g(t)^2}{2t}\right)$$

- (e) In the last part we established the exact density of the hitting time of a Brownian motion approaching a linear boundary. We now generalize this to nonlinear boundaries, but at the cost that now we only have a bound. Suppose that $f : [0, \infty) \rightarrow [0, \infty)$ is concave and differentiable and let $\lambda_t : \mathbb{R} \rightarrow \mathbb{R}$ be the tangent to f at t given by $\lambda_t(x) = f(t) + f'(t)(x - t)$. Let $\tau_f = \min\{t : B_t = f(t)\}$ and $q_f(t)$ be the density of τ_f . Show that

$$q_f(t) \leq q_{\lambda_t}(t).$$

- (f) Suppose that X_1, X_2, \dots is a sequence of independent standard Gaussian random variables. Show that

$$\mathbb{P}\left(\text{exists } t \leq \infty : \sum_{s=1}^t X_s \geq f(t)\right) \leq \int_0^n \frac{\lambda_t(0)}{\sqrt{2\pi t^3}} \exp\left(-\frac{f(t)^2}{2t}\right) dt.$$

- (g) Let $h : (0, \infty) \rightarrow (1, \infty)$ be a concave monotone increasing function such that $\sqrt{\log(h(a))}/h(a) \leq c/a$ for constant $c > 0$ and $f(t) = \sqrt{2t \log h(1/t\delta)} + t\Delta$. Show that

$$\mathbb{P}\left(\text{exists } t \leq \infty : \sum_{s=1}^t X_s \geq f(t)\right) \leq \frac{2c\delta}{\sqrt{\pi}\Delta^2}.$$

- (h) Show that $h(a) = 1 + (1+a)\sqrt{\log(1+a)}$ satisfies the requirements of the previous part with $c = 11/10$.
- (i) Use your results to modify MOSS for the case when the rewards are Gaussian. Compare the algorithms empirically.
- (j) Prove for your modified algorithm that

$$\limsup_{n \rightarrow \infty} \frac{R_n}{\log(n)} \leq \sum_{i:\Delta_i > 0} \frac{2}{\Delta_i}.$$



The above exercise has several challenging components and assumes prior knowledge of Brownian motion and its interpretation in terms of the heat equation. We recommend the book by [Lerche \[1986\]](#) as a nice reference on hitting times for Brownian motion against concave barriers. The equation you derived in part (d) is called the Bachelier-Levy formula and the technique for doing so is the method of images. The use of this theory in bandits was introduced by one of the authors [[Lattimore, 2018](#)], which readers might find useful when working through these questions.

9.5 In the last exercise you modified MOSS to show asymptotic optimality when the noise is Gaussian. This is also possible for subgaussian noise. Follow the advice in the notes of this chapter to adapt MOSS so that for all 1-subgaussian bandits it holds that

$$\limsup_{n \rightarrow \infty} \frac{R_n}{\log(n)} \leq \sum_{i: \Delta_i > 0} \frac{2}{\Delta_i},$$

while maintaining the property that $R_n \leq C\sqrt{Kn}$ for universal constant $C > 0$.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

10 The Upper Confidence Bound Algorithm: Bernoulli Noise (\dagger)

In previous chapters we assumed the noise of the rewards was σ -subgaussian for some known $\sigma > 0$. This has the advantage of simplicity and relative generality, but stronger assumptions are sometimes justified and often lead to stronger results. In this chapter we consider the case where the rewards are Bernoulli ($X_t \in \{0, 1\}$). This is a fundamental setting found in many applications. For example, in click-through prediction the user either clicks on the link or not and so the reward is either zero or one. A Bernoulli bandit is characterized by the mean payoffs for each arm, $\mu_1, \dots, \mu_K \in [0, 1]$ and the reward observed in round t is $X_t \sim \mathcal{B}(\mu_{A_t})$.

We saw in Chapter 5 that the Bernoulli distribution is 1/2-subgaussian regardless of its mean, which means that UCB and its variants would enjoy logarithmic regret guarantees. However, the additional knowledge that the rewards are Bernoulli is not being fully exploited by these algorithms. The reason is essentially that the variance of a Bernoulli random variable depends on its mean, and when the variance is small the empirical mean concentrates faster, a fact that should be used to make the confidence intervals smaller.

10.1 Concentration for sums of Bernoulli random variables

Again we divert our attention away from bandits towards the concentration of the empirical mean towards the true value for sums of Bernoulli random variables. First we need to define a concept from information theory called the **relative entropy** or **Kullback-Leibler divergence**, which is a measure of similarity between distributions that for now we specify to the Bernoulli case. We defer the intuition for this concept until Chapter 14 where we give an introduction to information theory and specifically relative entropy.

DEFINITION 10.1 (Relative entropy between Bernoulli distributions) Let $p, q \in [0, 1]$. Then the relative entropy between Bernoulli distributions with parameters p and q respectively is defined to be

$$d(p, q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q)),$$

where the singularities are defined by taking limits so for $q \in [0, 1]$, $d(0, q) = \log(1/(1 - q))$, $d(1, q) = \log(1/q)$, and $d(p, 0) = d(p, 1) = \infty$ for $p \in (0, 1)$.

Notice that $d(p, q) = 0$ if and only if $p = q$ and $d(p, q) \geq 0$ for all p and q . So $d(\cdot, \cdot)$ is almost a metric except it is not symmetric and does not satisfy the triangle inequality. Some authors call such functions **premetrics**, but the nomenclature has not been standardized. The following lemma gives some useful properties of the relative entropy.

LEMMA 10.1 *Let $p, q, \varepsilon \in [0, 1]$. The following hold:*

- (a) *The functions $d(\cdot, q)$ and $d(p, \cdot)$ are convex and have unique minimizers at q and p respectively.*
- (b) *$d(p, q) \geq 2(p - q)^2$ (**Pinsker's inequality**).*
- (c) *If $p \leq q - \varepsilon \leq q$, then $d(p, q - \varepsilon) \leq d(p, q) - d(q - \varepsilon, q) \leq d(p, q) - 2\varepsilon^2$.*

The first inequality in (c) is a specialized version of the Pythagorean inequality for Bregman divergences. This is not important here, but see Chapter 26 for more details.

Proof For (a): $d(\cdot, q)$ is the sum of the negative binary entropy function $h(p) = p \log p + (1 - p) \log(1 - p)$ and a linear function. The second derivative of h is $h''(p) = 1/p + 1/(1 - p)$, which is positive and hence h is convex. For fixed p the function $d(p, \cdot)$ is the sum of $h(p)$ and convex functions $p \log(1/q)$ and $(1 - p) \log(1/(1 - q))$. Hence $d(p, \cdot)$ is convex. The minimizer property follows because $d(p, q) > 0$ unless $p = q$ in which case $d(p, p) = d(q, q) = 0$. A more general version of (b) is given in Chapter 15. A proof of the simple version here follows by considering the function $g(x) = d(p, p + x) - 2x^2$, which obviously satisfies $g(0) = 0$. The proof is finished by showing that this is the unique minimizer of g over the interval $[-p, 1 - p]$. The details are left to Exercise 10.1. For (c) notice that

$$h(p) = d(p, q - \varepsilon) - d(p, q) = p \log \frac{q}{q - \varepsilon} + (1 - p) \log \frac{1 - q}{1 - q + \varepsilon}.$$

It is easy to see then that h is linear and increasing in its argument. Therefore, since $p \leq q - \varepsilon$,

$$h(p) \leq h(q - \varepsilon) = -d(q - \varepsilon, q)$$

as required for the first inequality of (c). The second inequality follows by using the result in (b). \square

The next lemma controls the concentration of the sample mean of a sequence of independent and identically distributed Bernoulli random variables.

LEMMA 10.2 (Chernoff's bound) *Let X_1, X_2, \dots, X_n be a sequence of independent random variables that are Bernoulli distributed with mean μ and let $\hat{\mu} = \frac{1}{n} \sum_{t=1}^n X_t$ be the sample mean. Then for $\varepsilon \in [0, 1 - \mu]$ it holds that*

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp(-nd(\mu + \varepsilon, \mu)) \tag{10.1}$$

and for $\varepsilon \in [0, \mu]$ it holds that

$$\mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp(-nd(\mu - \varepsilon, \mu)). \quad (10.2)$$

Proof We will again use Chernoff's method. Let $\lambda > 0$ be some constant to be chosen later. Then

$$\begin{aligned} \mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) &= \mathbb{P}\left(\exp\left(\lambda \sum_{t=1}^n (X_t - \mu)\right) \geq \exp(\lambda n \varepsilon)\right) \\ &\leq \frac{\mathbb{E}[\exp(\lambda \sum_{t=1}^n (X_t - \mu))]}{\exp(\lambda n \varepsilon)} \\ &= (\mu \exp(\lambda(1 - \mu - \varepsilon)) + (1 - \mu) \exp(-\lambda(\mu + \varepsilon)))^n. \end{aligned}$$

This expression is minimized by $\lambda = \log \frac{(\mu+\varepsilon)(1-\mu)}{\mu(1-\mu-\varepsilon)}$. Therefore

$$\begin{aligned} \mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) &\leq \left(\mu \left(\frac{(\mu+\varepsilon)(1-\mu)}{\mu(1-\mu-\varepsilon)}\right)^{1-\mu-\varepsilon} + (1-\mu) \left(\frac{(\mu+\varepsilon)(1-\mu)}{\mu(1-\mu-\varepsilon)}\right)^{-\mu-\varepsilon}\right)^n \\ &= \left(\frac{\mu}{\mu+\varepsilon} \left(\frac{(\mu+\varepsilon)(1-\mu)}{\mu(1-\mu-\varepsilon)}\right)^{1-\mu-\varepsilon}\right)^n \\ &= \exp(-nd(\mu + \varepsilon, \mu)). \end{aligned}$$

The bound on the left tail is proven identically. \square

Using Pinsker's inequality, it follows that $\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon), \mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp(-2n\varepsilon^2)$, which is the same as what can be obtained from Hoeffding's lemma (see (5.8)). Solving $\exp(-2n\varepsilon^2) = \delta$ we recover the usual $1 - \delta$ confidence upper bound. In fact, this cannot be improved when $\mu \approx 1/2$, but the Chernoff bound is much stronger for μ is close to either zero or one. Can we invert the Chernoff tail bound to get confidence intervals which get tighter automatically as μ (or $\hat{\mu}$) approaches zero or one? The following corollary shows how to do this.

COROLLARY 10.1 *Let $\mu, \hat{\mu}, n$ be as above. Then, for any $a \geq 0$,*

$$\mathbb{P}(d(\hat{\mu}, \mu) \geq a, \hat{\mu} \leq \mu) \leq \exp(-na), \quad (10.3)$$

and

$$\mathbb{P}(d(\hat{\mu}, \mu) \geq a, \hat{\mu} \geq \mu) \leq \exp(-na). \quad (10.4)$$

Furthermore, defining

$$U(a) = \max\{u \in [0, 1] : d(\hat{\mu}, u) \leq a\},$$

it follows that with probability $1 - \exp(-na)$, $\mu < U(a)$. Similarly, letting

$$L(a) = \min\{u \in [0, 1] : d(\hat{\mu}, u) \leq a\},$$

with probability $1 - \exp(-na)$, $\mu > L(a)$ holds.

Proof First, we prove (10.3). Note that $d(\cdot, \mu)$ is decreasing on $[0, \mu]$, and thus, for $0 \leq a \leq d(0, \mu)$, $\{d(\hat{\mu}, \mu) \geq a, \hat{\mu} \leq \mu\} = \{\hat{\mu} \leq \mu - x, \hat{\mu} \leq \mu\} = \{\hat{\mu} \leq \mu - x\}$, where x is the unique solution to $d(\mu - x, \mu) = a$ on $[0, \mu]$. Hence, by Eq. (10.2) of Lemma 10.2, $\mathbb{P}(d(\hat{\mu}, \mu) \geq a, \hat{\mu} \leq \mu) \leq \exp(-na)$. When $a \geq d(0, \mu)$, the inequality trivially holds. The proof of (10.4) is entirely analogous and hence is omitted. For the second part of the corollary fix a and let $U = U(a)$. First notice that $U \geq \hat{\mu}$ and $d(\hat{\mu}, \cdot)$ is strictly increasing on $[\hat{\mu}, 1]$. Hence, $\{\mu \geq U\} = \{\mu \geq U, \mu \geq \hat{\mu}\} = \{d(\hat{\mu}, \mu) \geq d(\hat{\mu}, U), \mu \geq \hat{\mu}\} = \{d(\hat{\mu}, \mu) \geq a, \mu \geq \hat{\mu}\}$, where the last equality follows by $d(\hat{\mu}, U) = a$, which holds by the definition of U . Taking probabilities and using the first part of the corollary shows that $\mathbb{P}(\mu \geq U) \leq \exp(-na)$. The statement concerning $L = L(a)$ follows with a similar reasoning. \square

Note that for $\delta \in (0, 1)$, $U = U(\log(1/\delta)/n)$ and $L = L(\log(1/\delta)/n)$ are, respectively, upper and lower confidence bounds for μ . While U and L are defined implicitly in terms of an optimization problem. Although the relative entropy has no closed form inverse, the optimization can be solved to a high degree of accuracy using Newton's method (the relative entropy d is convex in its second argument). The advantage of this confidence interval to the one based on Hoeffding's interval is now clear: As $\hat{\mu}$ approaches one, the width of the interval, $U(a) - \hat{\mu}$ approaches zero, whereas the width of our previous interval stays $\sqrt{\log(1/\delta)/(2n)}$, a constant. The same holds for $\hat{\mu} - L(a)$ as $\hat{\mu} \rightarrow 0$.

EXAMPLE 10.1 Fig. 10.1 shows a plot of $d(3/4, x)$ and the lower bound given by Pinsker's inequality. The approximation degrades as $|x - 3/4|$ grows large, especially for $x > 3/4$. As explained in Corollary 10.1, the graph of $d(\hat{\mu}, \cdot)$ can be used to derive confidence bounds by solving for $d(\hat{\mu}, x) = a = \log(1/\delta)/n$. Assuming $\hat{\mu} = 3/4$ is observed, a confidence level of 90% with $n = 10$, $a \approx 0.23$. The confidence interval ends can then be read out from the figure by finding those values where the horizontal dashed black line intersects the solid blue line. The resulting confidence interval will be highly asymmetric. Note that in this scenario the lower confidence bounds produced by both Hoeffding's inequality and Chernoff's bound are similar while the upper bound provided by Hoeffding's bound is vacuous.

10.2 The KL-UCB algorithm

The KL-UCB algorithm is nothing more than UCB, but with Chernoff's bound used to define the upper confidence bound, rather than Lemma 5.1.

THEOREM 10.1 *If the reward in round t is $X_t \sim \mathcal{B}(\mu_{A_t})$, then the regret of*

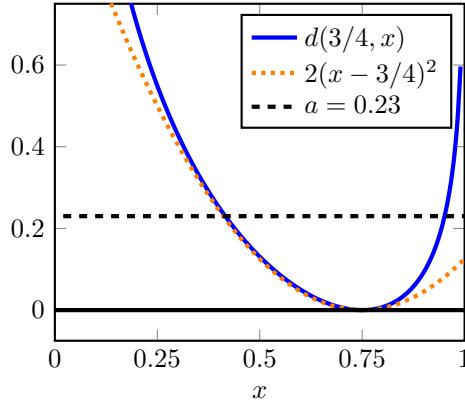


Figure 10.1 Relative entropy and Pinsker’s inequality

- 1: **Input** K
- 2: Choose each arm once
- 3: Subsequently choose

$$A_t = \operatorname{argmax}_i \max \left\{ \tilde{\mu} \in [0, 1] : d(\hat{\mu}_i(t-1), \tilde{\mu}) \leq \frac{\log f(t)}{T_i(t-1)} \right\},$$

where $f(t) = 1 + t \log^2(t)$.

Algorithm 7: KL-UCB

Algorithm 7 is bounded by

$$R_n \leq \sum_{i: \Delta_i > 0} \inf_{\varepsilon_1, \varepsilon_2 > 0, \varepsilon_1 + \varepsilon_2 \in (0, \Delta_i)} \Delta_i \left(\frac{f(n)}{d(\mu_i + \varepsilon_1, \mu^* - \varepsilon_2)} + \frac{1}{2\varepsilon_1^2} + 1 + \frac{1}{\varepsilon_2^2} \right).$$

$$\text{Furthermore, } \limsup_{n \rightarrow \infty} \frac{R_n}{\log(n)} \leq \sum_{i: \Delta_i > 0} \frac{\Delta_i}{d(\mu_i, \mu^*)}.$$

Let us now compare the asymptotic result above to that given for UCB in Theorem 8.1. Specializing this result for Bernoulli rewards (which are 1/2-subgaussian), we get

$$\limsup_{n \rightarrow \infty} \frac{R_n}{\log(n)} \leq \sum_{i: \Delta_i > 0} \frac{1}{2\Delta_i}.$$

By Pinsker’s inequality (part (b) of Lemma 10.1) we see that $d(\mu_i, \mu^*) \geq 2(\mu^* - \mu_i)^2 = 2\Delta_i^2$, which means that the asymptotic regret of KL-UCB is never worse than that of UCB. On the other hand, a Taylor’s expansion shows that when μ_i and μ^* are close (the hard case in the asymptotic regime), we have

$$d(\mu_i, \mu^*) = \frac{\Delta_i^2}{2\mu_i(1 - \mu_i)} + o(\Delta_i^2),$$

indicating that the regret of KL-UCB is approximately

$$\limsup_{n \rightarrow \infty} \frac{R_n}{\log(n)} \approx \sum_{i: \Delta_i > 0} \frac{2\mu_i(1 - \mu_i)}{\Delta_i}. \quad (10.5)$$

We might not be so surprised to notice that $\mu_i(1 - \mu_i)$ is the variance of a Bernoulli distribution with mean μ_i . Now $\mu_i(1 - \mu_i) \leq 1/4$, which shows that KL-UCB is never worse than the asymptotically optimal variant of UCB presented in the last chapter. But when μ_i is close to either zero or one, then KL-UCB is a big improvement.

The proof of Theorem 10.1 relies on two lemmas. The first is used to show that the index of the optimal arm is never too far below its true value, while the second shows that the index of any other arm is not often much larger than the same value. These results mirror those given for UCB, but things are complicated by the non-symmetric and hard-to-invert divergence function.

For the next results we define $d^+(p, q) = d(p, q)\mathbb{I}\{p < q\}$, $p, q \in [0, 1]$.

LEMMA 10.3 *Let X_1, X_2, \dots, X_n be independent Bernoulli random variables with mean $\mu \in [0, 1]$, $\varepsilon > 0$ and*

$$\tau = \min \left\{ t : \max_{1 \leq s \leq n} d^+(\hat{\mu}_s, \mu - \varepsilon) - \frac{\log f(t)}{s} \leq 0 \right\}.$$

Then, $\mathbb{E}[\tau] \leq \frac{2}{\varepsilon^2}$.

Proof We start with a high probability bound and then integrate to control the expectation:

$$\begin{aligned} \mathbb{P}(\tau > t) &\leq \mathbb{P} \left(\exists 1 \leq s \leq n : d^+(\hat{\mu}_s, \mu - \varepsilon) > \frac{\log f(t)}{s} \right) \\ &\leq \sum_{s=1}^n \mathbb{P} \left(d^+(\hat{\mu}_s, \mu - \varepsilon) > \frac{\log f(t)}{s} \right) \\ &= \sum_{s=1}^n \mathbb{P} \left(d(\hat{\mu}_s, \mu - \varepsilon) > \frac{\log f(t)}{s}, \hat{\mu}_s < \mu - \varepsilon \right) \\ &\leq \sum_{s=1}^n \mathbb{P} \left(d(\hat{\mu}_s, \mu) > \frac{\log f(t)}{s} + 2\varepsilon^2, \hat{\mu}_s < \mu \right) \quad (\text{(c) of Lemma 10.1}) \\ &\leq \sum_{s=1}^n \exp \left(-s \left(2\varepsilon^2 + \frac{\log f(t)}{s} \right) \right) \quad (\text{Eq. (10.3) of Corollary 10.1}) \\ &\leq \frac{1}{f(t)} \sum_{s=1}^n \exp(-2s\varepsilon^2) \\ &\leq \frac{1}{2f(t)\varepsilon^2}. \end{aligned}$$

To finish, we integrate the tail:

$$\mathbb{E}[\tau] \leq \int_0^\infty \mathbb{P}(\tau \geq t) dt \leq \frac{1}{2\varepsilon^2} \int_0^\infty \frac{dt}{f(t)} \leq \frac{2}{\varepsilon^2}. \quad \square$$

LEMMA 10.4 *Let X_1, X_2, \dots, X_n be independent Bernoulli random variables with mean μ . Further, let $\Delta > 0$, $a > 0$ and define*

$$\kappa = \sum_{s=1}^n \mathbb{I}\left\{d(\hat{\mu}_s, \mu + \Delta) \leq \frac{a}{s}\right\}.$$

$$\text{Then, } \mathbb{E}[\kappa] \leq \inf_{\varepsilon \in (0, \Delta)} \left(1 + \frac{a}{d(\mu + \varepsilon, \mu + \Delta)} + \frac{1}{2\varepsilon^2}\right).$$

Proof Let $\varepsilon \in (0, \Delta)$ and $u = a/d(\mu + \varepsilon, \mu + \Delta)$. Then

$$\begin{aligned} \mathbb{E}[\kappa] &= \sum_{s=1}^n \mathbb{P}\left(d(\hat{\mu}_s, \mu + \Delta) \leq \frac{a}{s}\right) \\ &\leq \sum_{s=1}^n \mathbb{P}\left(\hat{\mu}_s \geq \mu + \varepsilon \text{ or } d(\mu + \varepsilon, \mu + \Delta) \leq \frac{a}{s}\right) \\ &\quad (d(\cdot, \mu + \Delta) \text{ is decreasing on } [0, \mu + \Delta]) \\ &\leq u + \sum_{s=\lceil u \rceil}^n \mathbb{P}(\hat{\mu}_s \geq \mu + \varepsilon) \\ &\leq u + \sum_{s=1}^{\infty} \exp(-sd(\mu + \varepsilon, \mu)) \quad (\text{Lemma 10.2}) \\ &\leq 1 + \frac{a}{d(\mu + \varepsilon, \mu + \Delta)} + \frac{1}{d(\mu + \varepsilon, \mu)} \\ &\leq 1 + \frac{a}{d(\mu + \varepsilon, \mu + \Delta)} + \frac{1}{2\varepsilon^2} \quad (\text{Pinsker's inequality/Lemma 10.1(b)}) \end{aligned}$$

as required. \square

Proof of Theorem 10.1 As in other proofs we assume without loss of generality that $\mu_1 = \mu^*$ and bound $\mathbb{E}[T_i(n)]$ for suboptimal arms i . To this end, fix a suboptimal arm i and let $\varepsilon_1 + \varepsilon_2 \in (0, \Delta_i)$ with both ε_1 and ε_2 positive. Define

$$\begin{aligned} \tau &= \min \left\{ t : \max_{1 \leq s \leq n} d^+(\hat{\mu}_{1s}, \mu_1 - \varepsilon_2) - \frac{\log f(t)}{s} \leq 0 \right\}, \text{ and} \\ \kappa &= \sum_{s=1}^n \mathbb{I}\left\{d(\hat{\mu}_{is}, \mu_i + \Delta_i - \varepsilon_2) \leq \frac{\log f(n)}{s}\right\}. \end{aligned}$$

Then, by a similar reasoning as in the proof of Theorem 8.1,

$$\begin{aligned}
\mathbb{E}[T_i(n)] &= \mathbb{E} \left[\sum_{t=1}^n \mathbb{I}\{A_t = i\} \right] \\
&\leq \mathbb{E}[\tau] + \mathbb{E} \left[\sum_{t=\tau+1}^n \mathbb{I}\{A_t = i\} \right] \\
&\leq \mathbb{E}[\tau] + \mathbb{E} \left[\sum_{t=1}^n \mathbb{I} \left\{ A_t = i \text{ and } d(\hat{\mu}_{i,T_i(t-1)}, \mu_1 - \varepsilon_2) \leq \frac{\log f(t)}{T_i(t-1)} \right\} \right] \\
&\leq \mathbb{E}[\tau] + \mathbb{E}[\kappa] \\
&\leq 1 + \frac{2}{\varepsilon_2^2} + \frac{f(n)}{d(\mu_i + \varepsilon_1, \mu^* - \varepsilon_2)} + \frac{1}{2\varepsilon_1^2},
\end{aligned}$$

where the second inequality follows since by the definition of τ , if $t > \tau$, then the index of the optimal arm is at least as large as $\mu_1 - \varepsilon_2$. The third inequality follows from the definition of κ as in the proof of Theorem 8.1. The final inequality follows from Lemmas 10.3 and 10.4. The first claim of the theorem is completed by substituting the above into the standard regret decomposition

$$R_n = \sum_{i=1}^k \Delta_i \mathbb{E}[T_i(n)].$$

The asymptotic claim is left as an exercise. \square

10.3 Notes

- 1 The new concentration inequality (Lemma 10.2) actually holds more generally for any sequence of independent and identically distributed random variables X_1, X_2, \dots, X_n provided only that $X_t \in [0, 1]$ almost surely. Therefore all results in this section also hold if the assumption that the noise is Bernoulli is relaxed to the case where it is simply supported in $[0, 1]$ (or other bounded sets by shifting/scaling).
- 2 Expanding on the previous note, all that is required is a bound on the moment generating function for random variables X where $X \in [0, 1]$ almost surely. Garivier and Cappé [2011, Lemma 9] noted that $f(x) = \exp(\lambda x) - x(\exp(\lambda) - 1) - 1$ is negative on $[0, 1]$ and so

$$\mathbb{E}[\exp(\lambda X)] \leq \mathbb{E}[X(\exp(\lambda) - 1) + 1] = \mu \exp(\lambda) + 1 - \mu,$$

which is precisely the moment generating function of the Bernoulli distribution with mean μ . Then the remainder of the proof of Lemma 10.2 goes through unchanged. This shows that for any bandit $\nu = (P_i)_i$ with $\text{Supp}(P_i) \in [0, 1]$ for

all i the regret of the policy in Algorithm 7 satisfies

$$\limsup_{n \rightarrow \infty} \frac{R_n}{\log(n)} \leq \sum_{i: \Delta_i > 0} \frac{\Delta_i}{d(\mu_i, \mu^*)}.$$

- 3 The bounds obtained using the argument in the previous note are not quite tight. Specifically one can show there exists an algorithm such that for all bandits $\nu = (P_i)_i$ with P_i the reward distribution of the i th arm supported on $[0, 1]$, then

$$\limsup_{n \rightarrow \infty} \frac{R_n}{\log(n)} = \sum_{i: \Delta_i > 0} \frac{\Delta_i}{d_{[0,1]}(P_i, \mu^*)},$$

where

$$d_{[0,1]}(P_i, \mu^*) = \inf\{\text{D}(P_i, P) : \mu(P) > \mu^* \text{ and } \text{Supp}(P) \subset [0, 1]\}.$$

This last quantity is never smaller than $d(\mu_i, \mu^*)$. For details on this we refer the reader to the paper by [Honda and Takemura \[2010\]](#).

- 4 The approximation in Eq. (10.5) was used to show that the regret for KL-UCB is closely related to the variance of the Bernoulli distribution. It is natural to ask whether or not this result could be derived, at least asymptotically, by appealing to the central limit theorem. The answer is no! First, the quality of the approximation in Eq. (10.5) does not depend on n , so asymptotically it is not true that the Bernoulli bandit behaves like a Gaussian bandit with variances tuned to match. The reason is that as n tends to infinity, the confidence level should be chosen so that the risk of failure also tends to zero. But the central limit theorem does not provide information about the tails with probability mass less than $O(n^{-1/2})$. See Note 1 in Chapter 10.
- 5 This style of analysis is easily generalized to a wide range of alternative noise models, with the easiest being single parameter exponential families (Exercise 10.4).
- 6 Chernoff credits Lemma 10.2 to his friend Herman Rubin [[Chernoff, 2014](#)], but the name seems to have stuck.

10.4 Bibliographic remarks

Several authors have worked on Bernoulli bandits and the asymptotics have been well-understood since the article by [Lai and Robbins \[1985\]](#). The earliest version of the algorithm presented in this chapter is due to [Lai \[1987\]](#) who provided asymptotic analysis. The finite-time analysis of KL-UCB was given by two groups simultaneously (and published in the same conference!) by [Garivier and Cappé \[2011\]](#) and [Maillard et al. \[2011\]](#) (see also the combined journal article: [Cappé et al. 2013](#)). Two alternatives are the DMED [Honda and Takemura \[2010\]](#) and IMED [[Honda and Takemura, 2015](#)] algorithms. These works go after the problem of understanding the asymptotic regret for the more general

situation where the rewards lie in a bounded interval (see Note 3). The latter work covers even the semi-bounded case where the rewards are almost surely upper bounded. Both algorithms are asymptotically optimal. Ménard and Garivier [2017] combined MOSS and KL-UCB to derive an algorithm that is minimax optimal and asymptotically optimal for single parameter exponential families. While the subgaussian and Bernoulli examples are very fundamental, there has also been work on more generic setups where the unknown reward distribution for each arm is known to lie in some class \mathcal{F} . The article by Burnetas and Katehakis [1996] gives the most generic (albeit, asymptotic) results. These generic setups remain wide open for further work.

10.5 Exercises

10.1 [Pinsker's inequality] Prove Lemma 10.1(b).



Consider the function $g(x) = d(p, p + x) - 2x^2$ over the $[-p, 1 - p]$ interval. By taking derivatives, show that $g \geq 0$.

10.2 Let $\mathbb{F} = (\mathcal{F}_t)_t$ be a filtration, $(X_t)_t$ be $[0, 1]$ -valued, \mathbb{F} -adapted sequence, such that $\mathbb{E}[X_t | \mathcal{F}_{t-1}] = \mu_t$ for some $\mu_1, \dots, \mu_n \in [0, 1]$ non-random numbers. Define $\mu = \frac{1}{n} \sum_{t=1}^n \mu_t$, $\hat{\mu} = \frac{1}{n} \sum_{t=1}^n X_t$. Prove that the conclusion of Lemma 10.2 still holds.



Read note 2 at the end of this chapter. Let $g(\cdot, \mu)$ be the cumulant generating function of the μ -parameter Bernoulli distribution: For $X \sim \mathcal{B}(\mu)$, $\lambda \in \mathbb{R}$, $g(\lambda, \mu) = \log \mathbb{E}[\exp(\lambda X)]$. Show that $g(\lambda, \cdot)$ is concave. Next, use this and the tower rule to show that $\mathbb{E}[\exp(\lambda n(\hat{\mu} - \mu))] \leq g(\lambda, \mu)^n$.



The bound of the previous exercise is most useful when all μ_t are either all close to 0 or they are all close to 1. In particular, if half of the $\{\mu_t\}$ is close to zero, half of them is close to one, the bound will degrade to Hoeffding's bound. Irrespective of this, it is useful to notice that the claims made in Corollary 10.1 continue to hold for $\hat{\mu}, \mu$ as defined in the exercise.

10.3 Prove the asymptotic claim in Theorem 10.1.



Choose $\varepsilon_1, \varepsilon_2$ to decrease slowly with n and use the first part of the theorem.

10.4 Let h be a measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $T : \mathbb{R} \rightarrow \mathbb{R}$. The function T is called

the **sufficient statistic**. Define

$$\frac{dP_\theta}{dh}(x) = \exp(\theta T(x) - A(\theta)),$$

where $A(\theta)$ is the **log partition function** given by

$$A(\theta) = \log \int_{\mathbb{R}} \exp(\theta T(x)) dh(x).$$

Let $\Theta = \{\theta \in \mathbb{R} : A(\theta) \text{ exists}\}$. The set $\{P_\theta : \theta \in \Theta\}$ is called an **exponential family**. For more details see the note after the exercise.

- (a) Prove that for $\theta \in \Theta$ the function $P_\theta : \mathfrak{B}(\mathbb{R}) \rightarrow [0, 1]$ given by

$$P_\theta(A) = \int_A \frac{dP_\theta}{dh}(x) dh(x)$$

is a probability measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$.

- (b) Let \mathbb{E}_θ denote expectations with respect to P_θ and show that $A'(\theta) = \mathbb{E}_\theta[T(x)]$.
- (c) Find a choice of h and T such that $\{P_\theta : \theta \in \Theta\}$ is the family of Bernoulli distributions.
- (d) Find a choice of h and T such that $\{P_\theta : \theta \in \Theta\}$ is the family of Gaussian distributions with unit variance means in \mathbb{R} .
- (e) Let $\theta \in \Theta$ and $X \sim P_\theta$. Show that

$$\mathbb{E}_\theta[\exp(\lambda T(X))] = \exp(A(\lambda + \theta) - A(\theta)).$$

- (f) Given $\theta, \theta' \in \Theta$, show that

$$d(\theta, \theta') = \mathbb{E}_\theta \left[\log \left(\frac{p_\theta(X)}{p_{\theta'}(X)} \right) \right] = A(\theta') - A(\theta) - (\theta' - \theta) A'(\theta).$$

- (g) Let $\theta, \theta' \in \Theta$ be such that $A'(\theta') \geq A'(\theta)$ and X_1, \dots, X_n be independent and identically distributed and $\hat{T} = \frac{1}{n} \sum_{t=1}^n T(X_t)$. Show that

$$\mathbb{P} \left(\hat{T} \geq A'(\theta') \right) \leq \exp(-nd(\theta, \theta')) ,$$

- (h) Let \mathcal{E} be the set of all bandits with reward distributions in family $\{P_\theta : \theta \in \Theta\}$. Design a policy π such that for all $\nu \in \mathcal{E}$ it holds that

$$\lim_{n \rightarrow \infty} \frac{R_n(\pi, \nu)}{\log(n)} \leq \sum_{i: \Delta_i > 0} \frac{\Delta_i}{d(\theta_i, \theta_{i^*})},$$

where θ_i is such that P_{θ_i} is the distribution of the rewards for arm i and i^* is the optimal arm.



Exponential families represent a wide range of statistical models. We discuss them in more detail in Chapter 34. The function $d(\theta, \theta')$ is called the **relative entropy** between P_θ and $P_{\theta'}$. We discuss this concept more in Chapter 14. The bound in the last part of the exercise cannot be improved as we explain in Chapter 16.

10.5 In this exercise you compare KL-UCB and UCB empirically.

- (a) Implement Algorithm 7 and Algorithm 5 where the latter algorithm should be tuned for 1/2-subgaussian bandits so that

$$A_t = \operatorname{argmax}_{i \in [K]} \hat{\mu}_i(t-1) + \sqrt{\frac{\log(f(t))}{2T_i(t-1)}}.$$

- (b) Let $n = 10000$ and $K = 2$. Plot the expected regret of each algorithm as a function of Δ when $\mu_1 = 1/2$ and $\mu_2 = 1/2 + \Delta$.
- (c) Repeat the above experiment with $\mu_1 = 1/10$ and $\mu_2 = 9/10$.
- (d) Discuss your results.

Part III

Adversarial Bandits with Finitely Many Arms

Statistician George E. P. Box is famous for writing “All models are wrong, but some are useful”. In the stochastic bandit model we assumed the learner’s reward is generated at random from a distribution that depends only on the chosen action. It does not take much thought to realize this model is almost always wrong. At the macroscopic level typically considered in bandit problems there is not much that is stochastic about the world. And even if there was, it is hard to rule out the existence of other factors (observed or otherwise) influencing the rewards.

The quotation suggests we should not care whether or not the stochastic bandit model is right. Instead, we should ask if it is useful. In science, models are used for predicting the outcomes of future experiments and their usefulness is measured by the quality of the predictions. But how can this be applied to bandit problems? What predictions can be made based on bandit models? In this respect, we postulate the following.



The point of bandit models is to predict the performance of algorithms on future problem instances.

A model can fail in two fundamentally different ways. It can be too specific, imposing assumptions that are so detached from reality that a catastrophic mismatch between actual and predicted performance may arise.



Not all assumptions are equally important. It is a critical assumption in stochastic bandits that the mean reward of individual arms does not change (significantly) over time. On the other hand, the assumption that a single, arm-dependent distribution generates the rewards for a given arm plays a relatively insignificant role. The reader is encouraged to think of cases when the constancy of arm-distributions plays no role, and also of cases when it does. And furthermore, to decide to what extent the algorithms can tolerate deviations from the assumption that the means of arms stay the same. Stochastic bandits where the means of the arms are changing over time are called **nonstationary** and are the topic of Chapter 31.

The second mode of failure occurs when a model is too general, which makes the resulting algorithm overly cautious and harms performance.



If a highly specialized model is actually correct, then the resulting algorithms usually dominate algorithms derived for a more general model. This is a general manifestation of the bias-variance tradeoff, well known in supervised learning and statistics. The holy grail is to find algorithms that work ‘optimally’ across a range of models. The reader should think about examples from the previous chapters that illustrate these points.

The usefulness of the stochastic model depends on the setting. In particular, the designer of the bandit algorithm must carefully evaluate whether stochasticity, stability of the mean and independence are reasonable assumptions. For some applications the answer will probably be yes, while in others the practitioner may seek something more robust. This latter situation is the topic of the next few chapters.

Adversarial bandits

The **adversarial bandit** model abandons almost all the assumptions on how the rewards are generated, so much so that the ‘environment’ is now often called the ‘adversary’. The adversary has a great deal of power in this model, including the ability to examine the code of the proposed algorithms and choose the rewards accordingly. All that is kept from the previous chapters is that the objective will be framed in terms of how well an algorithm is able to compete with the best action in hindsight.

At first sight it seems remarkable that one can say anything at all about such a general model. And yet, it turns out that this model is not much harder than the stochastic bandit problem. Why this holds and how to design algorithms that achieve these guarantees will be explained in the following chapters.

To give you a glimmer of hope, imagine playing the following simple bandit game with a friend. The horizon is $n = 1$ and you have two actions. The game proceeds as follows:

- 1 You tell your friend your strategy for choosing an action.
- 2 Your friend secretly chooses rewards $x_1 \in \{0, 1\}$ and $x_2 \in \{0, 1\}$.
- 3 You implement your strategy to select $A \in \{1, 2\}$ and receive reward x_A .
- 4 The regret is $R = \max\{x_1, x_2\} - x_A$.

Clearly if your friend chooses $x_1 = x_2$, then your regret is zero no matter what. Now let's suppose you implement the deterministic strategy $A = 1$. Then your friend can choose $x_1 = 0$ and $x_2 = 1$ and your regret is $R = 1$. The trick to improve on this is to randomize. If you tell your friend: “I will choose $A = 1$ with probability one half”, then the best she can do is choose $x_1 = 1$ and $x_2 = 0$ (or reversed) and your expected regret is $R = 1/2$. You are forgiven if you did not settle on this solution yourself because we did not tell you that a strategy may be randomized. With such a short horizon you cannot do better than this, but for longer games the relative advantage of the adversary decreases.

Notes

- 1 Having derived a bandit algorithm one can ask how much it takes to break the guarantees. For example, under what circumstances is the regret of UCB

bounded by $C\sqrt{nK \log(n)}$? The lazy way is to push part of the proof into the assumptions. For UCB this might mean replacing a subgaussian assumption with a condition that the data generating processes satisfies the conclusion of the core concentration result (Corollary 5.1). A more ambitious goal is to define the subset of rewards for which the regret is bounded by some value and try to characterize this set. To our knowledge these ideas have not been explored in bandits and barely at all in machine learning more broadly.

Bibliographic remarks

The quote by George Box was used several times with different phrasings [Box, 1976, 1979]. The adversarial framework has its roots in game theory with familiar names like Hannan [1957] and Blackwell [1954] producing some of the early work. The nonstatistical approach has enjoyed enormous popularity since the 1990's and has been adopted wholeheartedly by the theoretical computer science community [Vovk, 1990, Littlestone and Warmuth, 1994, and many many others]. For bandits the earliest work that we know of is by Auer et al. [1995]. There is now a big literature on adversarial bandits, which we will cover in more depth in the chapters that follow.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

11 The Exp3 Algorithm

Let $K > 1$ be the number of arms. A **K -armed adversarial bandit** is an arbitrary sequence of reward vectors $\nu = (x_1, \dots, x_n)$ where $x_t \in [0, 1]^K$ for each $t \in [n]$. In each round the learner chooses an action $A_t \in [K]$ and observes reward $X_t = x_{tA_t}$. We do not capitalize the reward vectors (x_t) because they are not random. The learner will usually randomize their decisions so that A_t and X_t are random variables and hence capitalized.

Like in the stochastic setting, a policy can be viewed as a mapping from interaction sequences to a distribution over the actions. For stochastic bandits we did not yet make use a randomized policy, but for adversarial bandits this is crucial. Given a policy π the conditional distribution over the actions having observed $A_1, X_1, \dots, A_{t-1}, X_{t-1}$ is $P_t = \pi(\cdot | A_1, X_1, \dots, A_{t-1}, X_{t-1}) \in \mathcal{P}_{K-1}$.

The performance of a policy π on environment ν is measured by the expected regret, which is the expected loss in revenue of the policy relative to the best fixed action in hindsight.

$$R_n(\pi, \nu) = \max_i \sum_{t=1}^n x_{ti} - \mathbb{E} \left[\sum_{t=1}^n x_{tA_t} \right]. \quad (11.1)$$

When π and ν are clear from the context we may just write R_n in place of $R_n(\pi, \nu)$.

-  The only source of randomness in the regret comes from the randomness in the actions of the learner. Of course the interaction with the environment means the action chosen in round t may depend on actions $s < t$ as well as the observed rewards until round t .

Like in the stochastic setting, we are often interested in the worst-case regret over all environments, which is

$$R_n^*(\pi) = \sup_{\nu \in [0, 1]^{nK}} R_n(\pi, \nu).$$

The main question is whether or not there exist policies π for which $R_n^*(\pi)$ is sublinear in n . In Exercise 11.2 you will show that for deterministic policies $R_n^*(\pi) \geq n(1 - 1/K)$, which follows by constructing a bandit so that $x_{tA_t} = 0$ for

all t and $x_{ti} = 1$ for $i \neq A_t$. Because of this, sublinear worst-case regret is only possible by using a randomized policy.



Readers familiar with game theory will not be surprised by the need for randomization. The interaction between learner and adversarial bandit can be framed as a two-player zero-sum game between the environment and learner. The moves for the environment are the possible reward sequences and for the player they are the set of policies. The payoff for the environment/learner is the regret and its negation respectively. Since the player goes first, the only way to avoid being exploited is to choose a policy that randomizes.

While stochastic and adversarial bandits seem quite different, it turns out that the optimal worst case regret is the same up to constant factors and that lower bounds for adversarial bandits are invariably derived in the same manner as for stochastic bandits (see Part IV). In this chapter we present a simple algorithm for which the worst-case regret is suboptimal by just a logarithmic factor. First though, we explore the differences and similarities between stochastic and adversarial environments.

We already noted that deterministic strategies will have linear regret for some adversarial bandit. Since all the strategies in Part II were deterministic, they are not well suited for the adversarial setting. This immediately implies that policies that are good for stochastic bandit can be very suboptimal in the adversarial setting. What about the other direction? Will an adversarial bandit strategy have small expected regret in the stochastic setting? Let π be an adversarial bandit policy and $\nu = (\nu_1, \dots, \nu_K)$ be a stochastic bandit with ν_i supported on a subset of $[0, 1]$ for each i . Next let X_{ti} be sampled from ν_i for each $i \in [K]$ and $t \in [n]$ and assume these random variables are mutually independent. Then by Jensen's inequality and convexity of the maximum function we have

$$\begin{aligned} R_n(\nu, \pi) &= \max_i \mathbb{E} \left[\sum_{t=1}^n (X_{ti} - X_{tA_t}) \right] \leq \mathbb{E} \left[\max_i \sum_{t=1}^n (X_{ti} - X_{tA_t}) \right] \\ &= \mathbb{E} [R_n(\pi, (X_{ti}))] \leq R_n^*(\pi), \end{aligned}$$

where the regret in the first line is the stochastic regret and in the last it is the adversarial regret. Therefore the worst-case stochastic regret is upper bounded by the worst-case adversarial regret. Going the other way, the above inequality also implies the worst-case regret for adversarial problems is lower bounded by the worst-case regret on stochastic problems with rewards bounded in $[0, 1]$. In Chapter 15 we prove the worst-case regret for stochastic bandits is at least $c\sqrt{nK}$, where $c > 0$ is a universal constant. And so for the same universal constant the minimax regret for adversarial bandits satisfies

$$R_n^* = \inf_{\pi} \sup_{\nu \in [0, 1]^{nK}} R_n(\pi, \nu) \geq c\sqrt{nK}.$$

11.1 Importance-weighted estimators

A key ingredient of all adversarial bandit algorithms is a mechanism for estimating the reward of unplayed arms. Recall that P_t is the conditional distribution of the action played in round t and let P_{ti} denote the conditional probability that the policy chooses action $A_t = i$,

$$P_{ti} = \mathbb{P}(A_t = i \mid X_1, \dots, X_{t-1}, A_1, \dots, A_{t-1}),$$

In what follows we assume that $P_{ti} > 0$ almost surely, which is true for all policies considered in this chapter. Until you know how to do it, estimating the reward for all arms simultaneously using only P_t and the observed reward seems like a hopeless endeavor. The idea is to use the **importance-weighted estimator** given by

$$\hat{X}_{ti} = \frac{\mathbb{I}\{A_t = i\} X_t}{P_{ti}}. \quad (11.2)$$

One way to get a first impression about the quality of an estimator is to calculate its mean and variance. Is the mean of \hat{X}_{ti} close to x_{ti} ? Does \hat{X}_{ti} have a small variance? Let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid A_1, X_1, \dots, A_{t-1}, X_{t-1}]$ denote the conditional expectation given the history up to time t . Then the conditional expectation of \hat{X}_{ti} satisfies

$$\mathbb{E}_t[\hat{X}_{ti}] = x_{ti}, \quad (11.3)$$

which means that \hat{X}_{ti} is an unbiased estimate of x_{ti} given whatever history has been generated. To see why Eq. (11.3) holds, let $A_{ti} = \mathbb{I}\{A_t = i\}$ so that $X_t A_{ti} = x_{ti} A_{ti}$ and

$$\hat{X}_{ti} = \frac{A_{ti}}{P_{ti}} x_{ti}.$$

Now $\mathbb{E}_t[A_{ti}] = P_{ti}$ and since P_{ti} is a function of $A_1, X_1, \dots, A_{t-1}, X_{t-1}$, we get

$$\mathbb{E}_t[\hat{X}_{ti}] = \mathbb{E}_t\left[\frac{A_{ti}}{P_{ti}} x_{ti}\right] = \frac{x_{ti}}{P_{ti}} \mathbb{E}_t[A_{ti}] = \frac{x_{ti}}{P_{ti}} P_{ti} = x_{ti}.$$

By the tower rule for conditional expectation, (11.3) also implies that $\mathbb{E}[\hat{X}_{ti}] = \mathbb{E}[\mathbb{E}_t[\hat{X}_{ti}]] = x_{ti}$. For the variance we proceed in the same manner by considering the conditional variance $\mathbb{V}_t[\hat{X}_{ti}]$, which for arbitrary random variable U is

$$\mathbb{V}_t[U] = \mathbb{E}_t[(U - \mathbb{E}_t[U])^2].$$

So $\mathbb{V}_t[\hat{X}_{ti}]$ is a random variable that measures the variance of \hat{X}_{ti} conditioned on the past. Calculating the conditional variance using the definition of \hat{X}_{ti} and Eq. (11.3) shows that

$$\mathbb{V}_t[\hat{X}_{ti}] = \mathbb{E}_t[\hat{X}_{ti}^2] - x_{ti}^2 = \mathbb{E}_t\left[\frac{A_{ti}x_{ti}^2}{P_{ti}^2}\right] - x_{ti}^2 = \frac{x_{ti}^2(1 - P_{ti})}{P_{ti}}. \quad (11.4)$$

This can be extremely large when P_{ti} is small and x_{ti} is bounded away from zero. In the notes and exercises we shall see to what extent this can cause trouble. The estimator in (11.2) is the first that comes to mind, but there are alternatives. For example,

$$\hat{X}_{ti} = 1 - \frac{\mathbb{I}\{A_t = i\}}{P_{ti}}(1 - X_t). \quad (11.5)$$

This estimator is still unbiased. Rewriting the formula in terms of $y_{ti} = 1 - x_{ti}$ and $Y_t = 1 - X_t$ and $\hat{Y}_{ti} = 1 - \hat{X}_{ti}$ leads to

$$\hat{Y}_{ti} = \frac{\mathbb{I}\{A_t = i\}}{P_{ti}} Y_t.$$

This is the same as (11.2) except that Y_t has replaced X_t . The terms y_{ti} , Y_t and \hat{Y}_{ti} should be interpreted as **losses**. Had we started with losses to begin with then this would have been the estimator that first came to mind. For obvious reasons, the estimator in Eq. (11.5) is called the **loss-based importance-weighted estimator**. The conditional variance of this estimator is essentially the same as Eq. (11.4):

$$\mathbb{V}_t[\hat{X}_{ti}] = \mathbb{V}_t[\hat{Y}_{ti}] = y_{ti}^2 \frac{1 - P_{ti}}{P_{ti}}.$$

The only difference is that the variance now depends on y_{ti}^2 rather than x_{ti}^2 . Which is better then depends on the rewards for arm i , with smaller rewards suggesting the superiority of the first estimator and larger rewards (or small losses) suggesting the superiority of the second estimator. At this stage, one could be suspicious about the role of zero in this argument. Can we change the estimator (either one of them) so that it is more accurate for actions whose reward is close to some specific value v ? Of course! Just change the estimator so that v is subtracted from the observed reward (or loss), then use the importance sampling formula, and subsequently add back v . The problem is that the optimal value of v depends on the unknown quantity being estimated. Also note that the dependence of the variance on P_{ti} is the same for both estimators and since the rewards are bounded it is this term that usually contributes most significantly. In Exercise 11.4 we ask you to show that all unbiased estimators in this setting are importance-weighted estimators.



Although the two estimators seem quite similar, it should be noted that the first estimator in takes values in $[0, \infty)$ while the second takes values in $(-\infty, 1]$. Soon we will see that this difference has a big impact on the usefulness of these estimators when used in the Exp3 algorithm.

11.2 The Exp3 algorithm

The importance weighted estimator provides us with the means to estimate the reward. The next step is to choose the distribution over actions $P_t = (P_{ti})_i$. The simplest algorithm for adversarial bandits is called Exp3, which stands for “**E**xponential-weight algorithm for **E**xploration and **E**xploitation”. The reason for this name will become clear after the explanation of the algorithm. Let $\hat{S}_{ti} = \sum_{s=1}^t \hat{X}_{si}$ be the total estimated reward by the end of round t . It seems natural to choose the action-selection probabilities so that actions with larger estimated reward receive more weight. While there are many ways to map \hat{S}_{ti} into probabilities, a simple and popular choice is called **exponential weighting**, which for tuning parameter $\eta > 0$ sets

$$P_{ti} = \frac{\exp(\eta \hat{S}_{t-1,i})}{\sum_j \exp(\eta \hat{S}_{t-1,j})}. \quad (11.6)$$

The parameter η is called the **learning rate** and its role is to control how aggressively P_{ti} is pushed towards arms for which the estimated cumulative reward is highest. As $\eta \rightarrow \infty$, the probability mass in P_t quickly concentrates on $\text{argmax}_i \hat{S}_{t-1,i}$. It is here that the exploration/exploitation dilemma raises its head. If η is large, then the resulting policy will explore with low probability. But when P_{ti} is small, then the variance of the importance-weighted estimator is large and the estimates for these arms could be very poor. The consequence is the usual tension. Large η leads to overconfident policies and small η leads to excessive exploration.

There are many ways to set η , including allowing it to vary with time. In this chapter we restrict our attention to the simplest case by choosing η to depend only on the number of actions K and the horizon n . Since the algorithm depends on η this means that the horizon must be known in advance. This is relaxed in subsequent chapters.



For practical implementations it is useful to note that P_t can be calculated incrementally by

$$P_{t+1,i} = \frac{P_{ti} \exp(\eta \hat{X}_{ti})}{\sum_{j=1}^K P_{tj} \exp(\eta \hat{X}_{tj})}. \quad (11.7)$$

Computing the summation in the denominator can be numerically unstable because its terms can vary by several orders of magnitude. There are a variety of approaches for summing floats in a numerically stable way. One of the simplest is Kahan’s algorithm [Kahan, 1965]. An even better approach is to note that Eq. (11.7) does not change if all \hat{S}_{ti} are translated by some fixed

```

1: Input:  $n, K, \eta$ 
2: Set  $\hat{S}_{0i} = 0$  for all  $i$ 
3: for  $t = 1, \dots, n$  do
4:   Calculate the sampling distribution  $P_t$ :
```

$$P_{ti} = \frac{\exp(\eta \hat{S}_{t-1,i})}{\sum_{j=1}^K \exp(\eta \hat{S}_{t-1,j})}$$

```

5:   Sample  $A_t \sim P_t$  and observe reward  $X_t$ 
6:   Calculate  $\hat{S}_{ti}$ :
```

$$\hat{S}_{ti} = \hat{S}_{t-1,i} + 1 - \frac{\mathbb{I}\{A_t = i\} (1 - X_t)}{P_{ti}}$$

```
7: end for
```

Algorithm 8: Exp3

amount. Let $\tilde{S}_{ti} = \hat{S}_{ti} - \min_j \hat{S}_{tj}$ so that

$$P_{t+1,i} = \frac{\exp(\eta \tilde{S}_{ti})}{\sum_{j=1}^K \exp(\eta \tilde{S}_{tj})}.$$

Care is still required for the summation in the denominator, but the number of floating point multiplications has been reduced significantly.

11.3 Regret analysis

We are now ready to bound the expected regret of Exp3 (Algorithm 8).

THEOREM 11.1 *Let $\nu = (x_{ti}) \in [0, 1]^{nK}$ be an arbitrary adversarial bandit and π be the policy of Exp3 (Algorithm 8) with learning rate $\eta = \sqrt{\log(K)/(nK)}$. Then*

$$R_n(\pi, \nu) \leq 2\sqrt{nK \log(K)}.$$

Proof For any arm i define

$$R_{ni} = \sum_{t=1}^n x_{ti} - \mathbb{E} \left[\sum_{t=1}^n X_t \right],$$

which is the expected regret relative to using action i in all the rounds. The result will follow by bounding R_{ni} for all i , including the optimal arm. For the remainder of the proof, let i be some fixed arm. By the unbiasedness property of

\hat{X}_{ti} ,

$$\mathbb{E}[\hat{S}_{ni}] = \sum_{t=1}^n x_{ti} \quad \text{and also} \quad \mathbb{E}_t[X_t] = \sum_{i=1}^K P_{ti}x_{ti} = \sum_{i=1}^K P_{ti}\mathbb{E}_t[\hat{X}_{ti}].$$

The tower rule says that $\mathbb{E}[\mathbb{E}_t[X_t]] = \mathbb{E}[X_t]$, which together with the linearity of expectation and the above display means that

$$R_{ni} = \mathbb{E}[\hat{S}_{ni}] - \mathbb{E}\left[\sum_{t=1}^n \sum_{i=1}^K P_{ti}\hat{X}_{ti}\right] = \mathbb{E}[\hat{S}_{ni} - \hat{S}_n], \quad (11.8)$$

where the last equality serves as the definition of $\hat{S}_n = \sum_{t,i} P_{ti}\hat{X}_{ti}$. To bound the right hand side of Eq. (11.8) let

$$W_t = \sum_{j=1}^K \exp(\eta\hat{S}_{tj}).$$

By convention an empty sum is zero, which means that $S_{0j} = 0$ and $W_0 = K$. Then

$$\exp(\eta\hat{S}_{ni}) \leq \sum_{j=1}^K \exp(\eta\hat{S}_{nj}) = W_n = W_0 \frac{W_1}{W_0} \cdots \frac{W_n}{W_{n-1}} = K \prod_{t=1}^n \frac{W_t}{W_{t-1}}.$$

The ratio in the product can be rewritten in terms of P_t by

$$\frac{W_t}{W_{t-1}} = \sum_{j=1}^K \frac{\exp(\eta\hat{S}_{t-1,j})}{W_{t-1}} \exp(\eta\hat{X}_{tj}) = \sum_{j=1}^K P_{tj} \exp(\eta\hat{X}_{tj}). \quad (11.9)$$

We need the following facts:

$$\exp(x) \leq 1 + x + x^2 \text{ for all } x \leq 1 \quad \text{and} \quad 1 + x \leq \exp(x) \text{ for all } x \in \mathbb{R}.$$

Using these two inequalities leads to

$$\frac{W_t}{W_{t-1}} \leq 1 + \eta \sum_{j=1}^K P_{tj}\hat{X}_{tj} + \eta^2 \sum_{j=1}^K P_{tj}\hat{X}_{tj}^2 \leq \exp\left(\eta \sum_{j=1}^K P_{tj}\hat{X}_{tj} + \eta^2 \sum_{j=1}^K P_{tj}\hat{X}_{tj}^2\right).$$

Notice that this was only possible because \hat{X}_{tj} is defined by Eq. (11.5), which ensures that $\hat{X}_{tj} \leq 1$ and would not have been true had we used Eq. (11.2). Putting the inequalities together we get

$$\exp(\eta\hat{S}_{ni}) \leq K \exp\left(\eta\hat{S}_n + \eta^2 \sum_{t=1}^n \sum_{j=1}^K P_{tj}\hat{X}_{tj}^2\right).$$

Taking the logarithm of both sides, dividing by $\eta > 0$ and reordering gives

$$\hat{S}_{ni} - \hat{S}_n \leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^n \sum_{j=1}^K P_{tj}\hat{X}_{tj}^2. \quad (11.10)$$

As noted earlier, the expectation of the left-hand side is R_{ni} . The first term on

the right-hand side is a constant, which leaves us to bound the expectation of the second term. Letting $y_{tj} = 1 - x_{tj}$ and $Y_t = 1 - X_t$, then expanding the definition of \hat{X}_{tj}^2 leads to

$$\begin{aligned}\mathbb{E} \left[\sum_{t,j} P_{tj} \hat{X}_{tj}^2 \right] &= \mathbb{E} \left[\sum_{t=1}^n \sum_{j=1}^K P_{tj} \left(1 - \frac{\mathbb{I}\{A_t = j\} y_{tj}}{P_{tj}} \right)^2 \right] \\ &= \sum_{t=1}^n \mathbb{E} \left[\sum_{j=1}^K P_{tj} \left(1 - 2 \frac{\mathbb{I}\{A_t = j\} y_{tj}}{P_{tj}} + \frac{\mathbb{I}\{A_t = j\} y_{tj}^2}{P_{tj}^2} \right) \right] \\ &= \sum_{t=1}^n \mathbb{E} \left[1 - 2Y_t + \mathbb{E}_t \left[\sum_{j=1}^K \frac{\mathbb{I}\{A_t = j\} y_{tj}^2}{P_{tj}} \right] \right] \\ &= \sum_{t=1}^n \mathbb{E} \left[1 - 2Y_t + \sum_{j=1}^K y_{tj}^2 \right] \\ &= \sum_{t=1}^n \mathbb{E} \left[(1 - Y_t)^2 + \sum_{j \neq A_t} y_{tj}^2 \right] \\ &\leq nK.\end{aligned}$$

By substituting this into Eq. (11.10), we get

$$R_{ni} \leq \frac{\log(K)}{\eta} + \eta nK = 2\sqrt{nK \log(K)},$$

where the equality follows by substituting $\eta = \sqrt{\log(K)/(nK)}$, which was chosen to optimize this bound. \square

At the heart of the proof are the inequalities:

$$1 + x \leq \exp(x) \text{ for all } x \in \mathbb{R} \quad \text{and} \quad \exp(x) \leq 1 + x + x^2 \text{ for } x \leq 1.$$

Attentive readers will notice that the former of these inequalities is an ansatz derived from the first order Taylor expansion of $\exp(x)$ about $x = 0$. The latter, however, is not the second order Taylor expansion, which would be $1 + x + x^2/2$. The problem is that the second order Taylor series is not an upper bound on $\exp(x)$ for $x \leq 1$, but only for $x \leq 0$:

$$\exp(x) \leq 1 + x + \frac{1}{2}x^2 \text{ for all } x \leq 0. \tag{11.11}$$

But it is nearly an upper bound, and this can be exploited to improve the bound in Theorem 11.1. The mentioned upper and lower bounds on $\exp(x)$ are shown in Fig. 11.1, from which it is quite obvious that the new bound is significantly tighter when $x \leq 0$.

Let us now put Eq. (11.11) to use in proving the following improved version of Theorem 11.1 for which the regret is smaller by a factor of $\sqrt{2}$. This looks quite insignificant, but in relative terms shaves off approximately thirty percent

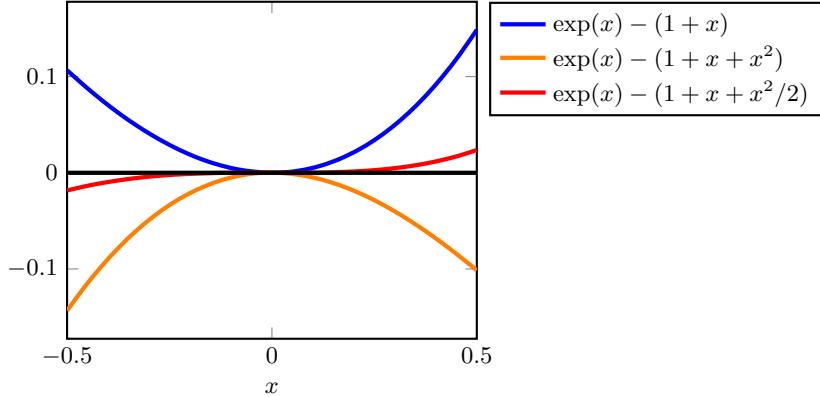


Figure 11.1 Approximations for $\exp(x)$ on $[-1/2, 1/2]$.

of the previous bound. The algorithm is unchanged except for a slightly increased learning rate.

THEOREM 11.2 *Let $\nu = (x_{ti}) \in [0, 1]^{nK}$ be an adversarial bandit and π be the policy of Exp3 with learning rate $\eta = \sqrt{2 \log(K)/(nK)}$. Then*

$$R_n(\pi, \nu) \leq \sqrt{2nK \log(K)}.$$

Proof By construction $\hat{X}_{tj} \leq 1$. Therefore

$$\begin{aligned} \exp(\eta \hat{X}_{tj}) &= \exp(\eta) \exp(\eta(\hat{X}_{tj} - 1)) \\ &\leq \exp(\eta) \left\{ 1 + \eta(\hat{X}_{tj} - 1) + \frac{\eta^2}{2}(\hat{X}_{tj} - 1)^2 \right\}. \end{aligned}$$

Using the fact that $\sum_j P_{tj} = 1$ and the inequality $1 + x \leq \exp(x)$ we get

$$\frac{W_t}{W_{t-1}} = \sum_{j=1}^K P_{tj} \exp(\eta \hat{X}_{tj}) \leq \exp \left(\eta \sum_{j=1}^K P_{tj} \hat{X}_{tj} + \frac{\eta^2}{2} \sum_{j=1}^K P_{tj} (\hat{X}_{tj} - 1)^2 \right),$$

where the equality is from Eq. (11.9). We see that here we need to bound $\sum_j P_{tj}(\hat{X}_{tj} - 1)^2$. Let $\hat{Y}_{tj} = 1 - \hat{X}_{tj}$. Then

$$P_{tj}(\hat{X}_{tj} - 1)^2 = P_{tj} \hat{Y}_{tj} \hat{Y}_{tj} = \mathbb{I}\{A_t = j\} y_{tj} \hat{Y}_{tj} \leq \hat{Y}_{tj},$$

where the last inequality used $\hat{Y}_{tj} \geq 0$ and $y_{tj} \leq 1$. Thus,

$$\sum_{j=1}^K P_{tj}(\hat{X}_{tj} - 1)^2 \leq \sum_{j=1}^K \hat{Y}_{tj}.$$

With the same calculations as before, we get

$$\hat{S}_{ni} - \hat{S}_n \leq \frac{\log(K)}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{j=1}^K \hat{Y}_{tj}. \quad (11.12)$$

The result is completed by taking expectations of both sides, using $\mathbb{E} \sum_{t,j} \hat{Y}_{tj} = \mathbb{E} \sum_{t,j} \mathbb{E}_t \hat{Y}_{tj} = \mathbb{E} \sum_{t,j} y_{tj} \leq nK$, and then substituting the learning rate. \square

11.4 Notes

- 1 The expected regret of Exp3 cannot be improved significantly, but the distribution of its regret is poorly behaved. Define the **random regret** to be the random variable measuring the actual deficit of the learner relative to the best arm in hindsight:

$$\hat{R}_n = \max_{i \in [K]} \underbrace{\sum_{t=1}^n x_{ti}}_{\text{in terms of rewards}} - \underbrace{\sum_{t=1}^n X_t}_{\sum_{t=1}^n Y_t - \min_{i \in [K]} \sum_{t=1}^n y_{ti}},$$

In Exercise 11.5 you will show that for all large enough n and reasonable choices of η there exists a bandit such that the random regret of Exp3 satisfies $\mathbb{P}(\hat{R}_n \geq n/4) > 1/131$. This is quite a troubling result and motivates the introduction of algorithms in the next chapter for which the distribution of \hat{R}_n is well behaved.

- 2 What happens when the range of the rewards is unbounded? This has been studied by [Allenberg et al. \[2006\]](#), where some (necessarily much weaker) positive results are presented.
- 3 A more basic problem than the one considered here is when the learner receives all $(x_{ti})_i$ at the end of round t , but the reward is still x_{tA_t} . This setting is called the **full-information** setting or **prediction with expert advice**. Exponential weighting is still a good idea, but the estimated rewards can now be replaced by the actual rewards. The resulting algorithm is sometimes called Hedge or the Exponential Weights Algorithm (EWA). The proof as written goes through in almost the same way, but one should replace the polynomial upper bound on $\exp(x)$ with Hoeffding's lemma. This analysis gives a regret of $\sqrt{n \log(K)/2}$, which is optimal in an asymptotic sense [[Cesa-Bianchi and Lugosi, 2006](#)].
- 4 A more sophisticated algorithm and analysis shaves a factor of $\sqrt{\log(K)}$ from the regret upper bound [[Audibert and Bubeck, 2009, 2010a, Bubeck and Cesa-Bianchi, 2012](#)]. The algorithm is an instantiation of the mirror descent algorithm from convex optimization, which we present in Chapter 28 for the more general adversarial linear bandit problem. Exercise 28.10 in that chapter explains the steps needed to solve this problem.
- 5 The initial distribution (the ‘prior’) P_1 does not have to be uniform. By biasing the prior towards a specific action the regret can be reduced when the favored action turns out to be optimal. There is a price for this, however, if the optimal arm is not favored [[Lattimore, 2015a](#)].
- 6 It was assumed in this chapter that the environment chose the rewards

at the start of the game. Such environments are called **oblivious** because the choices of the environment do not depend on those of the learner. A **reactive environment** is one where x_t is allowed to depend on the history $a_1, x_1, \dots, a_{t-1}, x_{t-1}$. Despite the fact that this is clearly a harder problem the result we obtained can be generalized to this setting without changes to the analysis. It is another question whether the definition of regret makes sense for such reactive environments.

- 7 Building on the previous note, suppose the reward vector in round t is $X_t = f_t(A_1, \dots, A_t)$ and f_1, \dots, f_n are a sequence of functions chosen in advance by the adversary with $f_t : [K]^t \rightarrow [0, 1]$. Let $\Pi \subset [K]^n$ be a set of action-sequences. Then the expected **policy regret** with respect to Π is

$$\max_{a_1, \dots, a_n \in \Pi} \sum_{t=1}^n f_t(a_1, \dots, a_t) - \mathbb{E} \left[\sum_{t=1}^n f_t(A_1, \dots, A_t) \right].$$

Even if Π only consists of constant sequences, there still does not exist a policy guaranteeing sublinear regret. The reason is simple. Consider the two candidate choices of f_1, \dots, f_n . In the first choice $f_t(a_1, \dots, a_t) = \mathbb{I}\{a_1 = 1\}$ and in the second we have $f_t(a_1, \dots, a_t) = \mathbb{I}\{a_1 = 2\}$. Clearly the learner must suffer linear regret in at least one of these two reactive bandit environments. The problem is that the learner's decision in the first round determines the rewards available in all subsequent rounds and there is no time for learning. By making additional assumptions sublinear regret is possible, however. For example, by assuming the adversary has limited memory [Arora et al., 2012].

- 8 There is a common misconception that the adversarial framework is a good fit for nonstationary environments. While the framework does not assume the rewards are stationary, the regret concept used in this chapter has stationarity built in. An algorithm that keeps the regret (as defined here) small is unsuitable for nonstationary environments because the best single action in hindsight is seldom a good benchmark when the environment is changing over time. Hence, the goal should be to compete with the sequence of actions in hindsight. For more on nonstationary bandits see Chapter 31.

- 9 The estimators in Eq. (11.2) and Eq. (11.5) both have conditional variance $\mathbb{V}_t[\hat{X}_{ti}] \approx 1/P_{ti}$, which blows up for small P_{ti} . It is instructive to think about whether and how P_{ti} can take on very small values. Consider the loss-based estimator given by (11.5). For this estimator, when P_{tA_t} and X_t are both small, \hat{X}_{tA_t} can take on a large negative value. Through the update formula (11.6) this then translates into P_{t+1,A_t} being squashed aggressively towards zero. A similar issue arises with the reward-based estimator given by (11.2). The difference is that now it will be a 'positive surprise' (P_{tA_t} small, X_t large) that pushes the probabilities towards zero. But note that in this case $P_{t+1,i}$ is pushed towards zero for all $i \neq A_t$. This means that dangerously small probabilities are expected to be more frequent for the gains estimator Eq. (11.2).

- 10 We argued at the beginning of the chapter that deterministic policies are

no good for adversarial bandit problems, which rules out all of the policies analyzed in Part II. We also showed the regret of Exp3 grows with at most the square root of the horizon on both stochastic and nonstochastic bandits. One might wonder if there exists a policy with (near-)optimal regret for adversarial bandits and logarithmic regret for stochastic bandits. There is a line of work addressing this question, which shows that such algorithm do exist [Bubeck and Slivkins, 2012, Seldin and Slivkins, 2014, Auer and Chiang, 2016, Seldin and Lugosi, 2017]. There are some complications, however, depending on whether or not the adversary is oblivious or not. The situation is best summarized by Auer and Chiang [2016], where the authors present upper and lower bounds on what is possible in various scenarios.

- 11 Exp3 requires advance knowledge of the horizon. The doubling trick can be used to overcome this issue, but perhaps a more elegant solution is to use a decreasing learning rate. The analysis in this chapter can be adapted to this case. More discussion is provided in the notes and exercises of Chapter 28 where we give a more generic solution to this problem.
- 12 There is a connection between adversarial learning and simultaneous-action zero sum games. This is discussed in a little more detail in the notes and exercises of Chapter 28.

11.5 Bibliographic remarks

Exponential weighting has been a standard tool in online learning since the papers by Vovk [1990] and Littlestone and Warmuth [1994]. Exp3 and several variations were introduced by Auer et al. [1995], which was also the first paper to study bandits in the adversarial framework. The algorithm and analysis presented here differs slightly because we do not add any additional exploration, while the version of Exp3 in that paper explores uniformly with low probability. The fact that additional exploration is not required was observed by Stoltz [2005].

11.6 Exercises

- 11.1** In order to implement Exp3 you need a way to sample from the exponential weights distribution. Many programming language provide a standard way to do this. For example in Python you can use the Numpy library and `numpy.random.multinomial`. In more basic languages, however, you only have access to a function `rand()` that returns a floating point number ‘uniformly’ distributed in $[0, 1]$. Describe an algorithm that takes as input a probability vector $p \in \mathcal{P}_{d-1}$ and uses a single call to `rand()` to return $X \in [d]$ with $\mathbb{P}(X = i) = p_i$.



Of course, on most computers `rand()` will return a pseudo-random number and since there are only finitely many floating point numbers the resulting distribution will not really be uniform on $[0, 1]$. Thinking about these issues is a worthy endeavour, and sometimes it really matters. For this exercise you may ignore these issues, however.

11.2 Show that for any deterministic policy π there exists an environment ν such that $R_n(\pi, \nu) \geq n(1 - 1/K)$. What does your result say about the policies design in Part II?

11.3 Suppose we had defined the regret by

$$R_n^{\text{track}}(\pi, \nu) = \mathbb{E} \left[\sum_{t=1}^n \max_{i \in [K]} x_{ti} - \sum_{t=1}^n x_{tA_t} \right].$$

At first sight this definition seems like the right thing because it measures what you actually care about. Unfortunately, however, it gives the adversary too much power. Show that for any policy π (randomised or not) there exists a $\nu \in [0, 1]^{Kn}$ such that

$$R_n^{\text{track}}(\pi, \nu) \geq n \left(1 - \frac{1}{K} \right).$$

11.4 Let $P \in \mathcal{P}_{K-1}$ be a probability vector and suppose $\hat{X} : [K] \times \mathbb{R} \rightarrow \mathbb{R}$ is a function such that for all $x \in \mathbb{R}^K$,

$$\mathbb{E}[\hat{X}(i, x_i)] = \sum_{i=1}^K P_i \hat{X}(i, x_i) = x_1.$$

Show there exists an $a \in \mathbb{R}$ such that $\hat{X}(i, x) = a + \frac{\mathbb{I}\{i=1\} x_1 - a}{P_1}$.

11.5 In this exercise you will show that if $\eta \in [n^{-p}, 1]$ for some $p \in (0, 1)$, then for sufficiently large n there exists bandits on which Exp3 has a constant probability of suffering linear regret and hence the variance of its regret is $\Omega(n^2)$. Let $x \in [1/4, 1/2]$ be a constant to be tuned subsequently and define two-armed adversarial bandit in terms of its losses by

$$y_{t1} = \begin{cases} 0 & \text{if } t \leq n/2 \\ 1 & \text{otherwise} \end{cases} \quad \text{and} \quad y_{t2} = \begin{cases} x & \text{if } t \leq n/2 \\ 0 & \text{otherwise.} \end{cases}$$

We are interested in analyzing the algorithm that samples $A_t \sim P_t$ where $P_{ti} \propto \exp(-\eta \sum_{s=1}^{t-1} \hat{Y}_{si})$ with $\hat{Y}_{si} = y_{si} A_{si} / P_{si}$.

(a) Define sequence of real-valued functions q_1, \dots, q_n on domain $[1/4, 1/2]$ inductively by $q_0(x) = 1/2$ and

$$q_{s+1}(x) = \frac{q_s(x) \exp(-\eta x / q_s(x))}{1 - q_s(x) + q_s(x) \exp(-\eta x / q_s(x))}.$$

Show for $t \leq 1 + n/2$ that $P_{t2} = q_{T_2(t-1)}(x)$.

- (b) Show that q_s is continuous on its domain and that $\frac{d}{dx}q_s(x) \leq 0$ for all $s \geq 0$.
- (c) Let $s = \min\{u : q_u(1/2) < 1/(8n)\}$. Show that for $s \geq 4$ there exists an $x \in [1/4, 1/2]$ such that $q_s(x) = 1/(8n)$.
- (d) Let s and x be as in the previous part. Show for large enough n it holds that $\sum_{u=1}^{s-1} 1/q_u(x) \leq n/8$.
- (e) Let $N(t)$ be a discrete counting process with $N(1) = 0$ and $N(t+1) - N(t) \in \{0, 1\}$ almost surely and $\mathbb{P}(N(t+1) - N(t) = 1 | N(t)) = q_{N(t)}(x)$. Prove that

$$\mathbb{P}\left(X \left(2 \sum_{u=1}^s 1/q_u(x)\right) \geq s\right) \geq \frac{1}{2}.$$

- (f) Prove that $\mathbb{P}(T_2(n/4) \geq s) \geq \frac{1}{2}$.
- (g) Let E be the event that $\sum_{t=1}^{n/2} \hat{Y}_{t2} \geq 2n$. Prove that

$$\mathbb{P}\left(\sum_{t=n/2+1}^n \mathbb{I}\{A_t = 1\} \geq n/2 \mid E\right) \geq 1 - n \exp(-\eta n).$$

- (h) Prove that $\mathbb{P}(E) \geq \frac{1}{2}(1 - \exp(-1/32))$.
- (i) Prove that $\mathbb{P}(\hat{R}_n \geq \frac{n}{4}) \geq \frac{1}{2}(1 - \exp(-1/32))(1 - n \exp(-\eta n))$.
- (j) You have shown that for large enough n , $\mathbb{P}(\hat{R}_n \geq cn) \geq c$ for some universal constant c . Explain why does this not contradict the proof that $R_n = \mathbb{E}[\hat{R}_n] = O(\sqrt{n})$.
- (k) Let $n = 10^5$ and $\eta = \sqrt{2 \log(2)/(2n)}$. Find the value of x satisfying the conditions in Part (c) and simulate Exp3 to demonstrate linear regret with constant probability.

11.6 Show that Theorem 11.1 stays valid for an adversarially stopped Exp3. That is, imagine that an adversary is given the power to stop Exp3 at some random time $\tau \in [n]$. The adversary is restricted in this decision in that while it can use A_1, \dots, A_t when deciding about whether Exp3 should be stopped in t , it cannot use A_{t+1}, \dots, A_n . That is, $\{\tau = t\}$ must be $\mathcal{F}_t = \sigma(A_1, \dots, A_t)$ -measurable. Show that $\mathbb{E}[\hat{R}_\tau] \leq 2\sqrt{nK \log(K)}$, where $\hat{R}_n = \sum_{t=1}^n x_{ti} - \sum_{t=1}^n X_t$ is the random regret of Exp3.



Use the identity $\sum_{t=1}^\tau U_t = \sum_{t=1}^n \mathbb{I}\{t \leq \tau\} U_t$, the tower rule and argue that $\{\tau \leq \tau\}$ is \mathcal{F}_{t-1} -measurable.

11.7 Let a_1, \dots, a_K be positive real values and U_1, \dots, U_K be a sequence of independent and identically distributed uniform random variables. Then let $G_i = -\log(-\log(U_i))$, which follows a **standard Gumbel distribution**. Prove

that

$$\mathbb{P} \left(\log(a_i) + G_i = \max_{k \in [K]} (\log(a_k) + G_k) \right) = \frac{a_i}{\sum_{k=1}^K a_k}.$$

 Let $(Z_{ti})_{ti}$ be a collection of independent and identically distributed random variables. The follow the perturbed leader algorithm chooses

$$A_t = \operatorname{argmax}_{i \in [K]} \left(Z_{ti} - \eta \sum_{s=1}^{t-1} \hat{\ell}_{ti} \right).$$

The previous exercise shows that choosing the distribution of Z_{ti} to be a standard Gumbel distribution makes follow the perturbed leader the same as exponential weights. This viewpoint will prove extremely useful when we tackle combinatorial bandits in Chapter 30.

11.8 In this exercise we compare UCB and Exp3 on stochastic data. Suppose we have a two-armed stochastic Bernoulli bandit with $\mu_1 = 0.5$ and $\mu_2 = \mu_1 + \Delta$ with $\Delta = 0.05$.

- (a) Plot the regret of UCB and Exp3 on the same plot as a function of the horizon n using the learning rate from Theorem 11.2.
- (b) Now fix the horizon to $n = 10^5$ and plot the regret as a function of the learning rate. Your plot should look like Fig. 11.2.
- (c) Investigate how the shape of this graph changes as you change Δ .
- (d) Find empirically the choice of η that minimizes the worst-case regret over all reasonable choices of Δ and compare to the value proposed by the theory.
- (e) What can you conclude from all this? Tell an interesting story.



The performance of UCB depends greatly on which version you use. For best results remember that Bernoulli distributions are 1/2-subgaussian or use the KL-UCB algorithm from Chapter 10.

11.9 Stress test your implementation of Exp3 from the previous exercise. What happens when $K = 2$ and the sequence of rewards is $x_{t1} = \mathbb{I}\{t \leq n/4\}$ and $x_{t2} = \mathbb{I}\{t > n/4\}$?

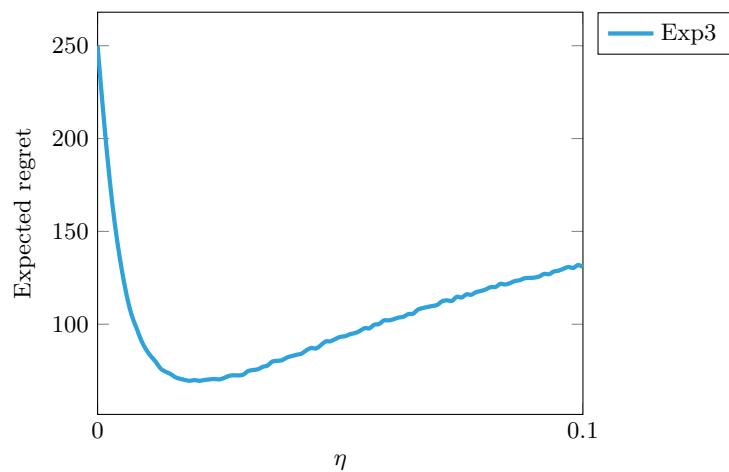


Figure 11.2 Expected regret for Exp3 for different learning rates over $n = 10^5$ rounds on a Bernoulli bandit with means $\mu_1 = 0.5$ and $\mu_2 = 0.55$.

12 The Exp3-IX Algorithm

In the last chapter we proved a sublinear bound on the expected regret of Exp3, but with a dishearteningly large variance. The objective of this chapter is to modify Exp3 so that the regret stays small in expectation and is simultaneously well concentrated about its mean. Such results are called **high probability bounds**.

One way to make Exp3 more robust is to make sure that P_{ti} is never too small. The first thing that comes to mind is to mix P_t with the uniform distribution. This is an explicit way of forcing exploration, which after further modification can be made to work. The resulting algorithm is called Exp3.P and we ask you to analyze it in Exercise 12.1. In this chapter we explore a similar idea that leads to an algorithm that is both simpler and empirically superior. The idea is to change the reward estimates to control the variance at the price of introducing some bias.

We start by summarizing what we know about the behaviour of the random regret of Exp3. Because we want to use the loss-based estimator it is more convenient to switch to losses, which we do for the remainder of the chapter. Rewriting Eq. (11.12) in terms of losses,

$$\hat{L}_n - \hat{L}_{ni} \leq \frac{\log(K)}{\eta} + \frac{\eta}{2} \sum_{j=1}^K \hat{L}_{nj}, \quad (12.1)$$

where \hat{L}_n and \hat{L}_{ni} are defined using the loss estimator \hat{Y}_{tj} by

$$\hat{L}_n = \sum_{t=1}^n \sum_{j=1}^K P_{tj} \hat{Y}_{tj} \quad \text{and} \quad \hat{L}_{ni} = \sum_{t=1}^n \hat{Y}_{ti}.$$



Eq. (12.1) holds no matter how the loss estimators are chosen provided they satisfy $\hat{Y}_{ti} \geq 0$ for all t and i . Of course the left-hand side of Eq. (12.1) is not close to the regret unless \hat{Y}_{ti} is a reasonable estimator of the loss y_{ti} ,

We also need to define the sum of losses observed by the learner and for each fixed action, which are

$$\tilde{L}_n = \sum_{t=1}^n y_{tA_t} \quad \text{and} \quad L_{ni} = \sum_{t=1}^n y_{ti}$$

Like in the previous chapter we need to define the (random) regret with respect to a given arm i as follows:

$$\hat{R}_{ni} = \sum_{t=1}^n x_{ti} - \sum_{t=1}^n X_t = \tilde{L}_n - L_{ni}. \quad (12.2)$$

By substituting the above definitions into Eq. (12.1) and rearranging the regret with respect to any arm i is bounded by

$$\begin{aligned} \hat{R}_{ni} &= \tilde{L}_n - L_{ni} = (\tilde{L}_n - \hat{L}_n) + (\hat{L}_n - \hat{L}_{ni}) + (\hat{L}_{ni} - L_{ni}) \\ &\leq \frac{\log(K)}{\eta} + (\tilde{L}_n - \hat{L}_n) + (\hat{L}_{ni} - L_{ni}) + \frac{\eta}{2} \sum_{j=1}^K \hat{L}_{nj}. \end{aligned} \quad (12.3)$$

This means the random regret can be bounded by controlling $\tilde{L}_n - \hat{L}_n$ and $\hat{L}_{nj} - L_{nj}$ and \hat{L}_{ni} . As promised we now modify the loss estimate. Let $\gamma > 0$ be a small constant to be chosen later and define the biased estimator

$$\hat{Y}_{ti} = \frac{\mathbb{I}\{A_t = i\} Y_t}{P_{ti} + \gamma}. \quad (12.4)$$

As γ increases the predictable variance decreases, but the bias increases. The optimal choice of γ depends on finding the sweet spot, which we will do once the dust has settled in the analysis. When Eq. (12.4) is used in the exponential update in Exp3, the resulting algorithm is called **Exp3-IX** (Algorithm 9). The suffix ‘IX’ stands for **i**mplicit exploration, a name justified by the following argument. A simple calculation shows that

$$\mathbb{E}_t[\hat{Y}_{ti}] = \frac{P_{ti}y_{ti}}{P_{ti} + \gamma} = y_{ti} - \frac{\gamma y_{ti}}{P_{ti} + \gamma} \leq y_{ti}.$$

Since small losses correspond to large rewards, the estimator is optimistically biased. The effect is a smoothing of P_t so that actions with large losses for which Exp3 would assign negligible probability are still chosen occasionally. As a result, Exp3-IX will explore more than the standard Exp3 algorithm (see Exercise 12.3). The reason for calling the exploration implicit is that it is a consequence of modifying the loss estimates, rather than directly altering P_t . This approach is more elegant mathematically and has nicer properties than the version that mixes P_t with the uniform distribution.

12.1 Regret analysis

We now prove the following theorem bounding the random regret of Exp3-IX with high probability.

THEOREM 12.1 *Let $\delta \in (0, 1)$ and define*

$$\eta_1 = \sqrt{\frac{2 \log(K+1)}{nK}} \quad \text{and} \quad \eta_2 = \sqrt{\frac{\log(K) + \log(\frac{K+1}{\delta})}{nK}}.$$

```

1: Input:  $n, K, \eta, \gamma$ 
2: Set  $\hat{L}_{0i} = 0$  for all  $i$ 
3: for  $t = 1, \dots, n$  do
4:   Calculate the sampling distribution  $P_t$ :

$$P_{ti} = \frac{\exp(-\eta \hat{L}_{t-1,i})}{\sum_{j=1}^K \exp(-\eta \hat{L}_{t-1,j})}$$

5:   Sample  $A_t \sim P_t$  and observe reward  $X_t$ 
6:   Calculate  $\hat{L}_{ti} = \hat{L}_{t-1,i} + \frac{\mathbb{I}\{A_t = i\}(1 - X_t)}{P_{t-1,i} + \gamma}$ 
7: end for

```

Algorithm 9: Exp3-IX

The following hold:

1 If Exp3-IX is run with parameters $\eta = \eta_1$ and $\gamma = \eta/2$, then

$$\mathbb{P}\left(\hat{R}_n \geq \sqrt{8.5nK \log(K+1)} + \left(\sqrt{\frac{nK}{2 \log(K+1)}} + 1\right) \log(1/\delta)\right) \leq \delta. \quad (12.5)$$

2 If Exp3-IX is run with parameters $\eta = \eta_2$ and $\gamma = \eta/2$, then

$$\mathbb{P}\left(\hat{R}_n \geq 2\sqrt{(2 \log(K+1) + \log(1/\delta))nK} + \log\left(\frac{K+1}{\delta}\right)\right) \leq \delta. \quad (12.6)$$



The value of η_1 is independent of δ , which means that using this choice of learning rate leads to a single algorithm with a high probability bound for all δ . On the other hand, η_2 does depend on δ so the user must choose a confidence level from the beginning. The advantage is that the bound is improved, but only for the specified confidence level. We will show in Chapter 17 that this tradeoff is unavoidable.

The proof follows by bounding each the terms in Eq. (12.3), which we do via a series of lemmas. The first of these lemmas is a new concentration bound, the statement of which requires us to introduce the notion of adapted and predictable sequences of random variables.

LEMMA 12.1 *Let $\mathbb{F} = (\mathcal{F}_t)_{0 \leq t \leq n}$ be a filtration and for $i \in [K]$ let $(\tilde{Y}_{ti})_t$ be \mathbb{F} -adapted such that:*

- 1 *For any $S \subset [K]$ with $|S| > 2$, $\mathbb{E}[\prod_{i \in S} \tilde{Y}_{ti} | \mathcal{F}_{t-1}] \leq 0$.*
- 2 *$\mathbb{E}[\tilde{Y}_{ti} | \mathcal{F}_{t-1}] = y_{ti}$ for all $t \in [n]$ and $i \in [K]$.*

Furthermore, let $(\alpha_{ti})_{ti}$ and $(\lambda_{ti})_{ti}$ be real-valued \mathcal{F}_t -predictable random sequences such that for all t, i it holds that $0 \leq \alpha_{ti} \tilde{Y}_{ti} \leq 2\lambda_{ti}$. Then for all $\delta \in (0, 1)$,

$$\mathbb{P}\left(\sum_{t=1}^n \sum_{i=1}^K \alpha_{ti} \left(\frac{\tilde{Y}_{ti}}{1 + \lambda_{ti}} - y_{ti}\right) \geq \log\left(\frac{1}{\delta}\right)\right) \leq \delta.$$

The proof relies on Chernoff's method and is deferred until the end of the chapter. Equipped with this result we can easily bound the terms $\hat{L}_{ni} - L_{ni}$.

LEMMA 12.2 *Let $\delta \in (0, 1)$. With probability at least $1 - \delta$ the following inequalities hold simultaneously:*

$$\max_{i \in [K]} (\hat{L}_{ni} - L_{ni}) \leq \frac{\log(\frac{K+1}{\delta})}{2\gamma} \quad \text{and} \quad \sum_{i=1}^K (\hat{L}_{ni} - L_{ni}) \leq \frac{\log(\frac{K+1}{\delta})}{2\gamma}. \quad (12.7)$$

Proof Fix $\delta' \in (0, 1)$ to be chosen later. Then

$$\sum_{i=1}^K (\hat{L}_{ni} - L_{ni}) = \sum_{t,i} \left(\frac{A_{ti} y_{ti}}{P_{ti} + \gamma} - y_{ti} \right) = \frac{1}{2\gamma} \sum_{t,i} 2\gamma \left(\frac{1}{1 + \frac{\gamma}{P_{ti}}} \frac{A_{ti} y_{ti}}{P_{ti}} - y_{ti} \right).$$

Introduce $\lambda_{ti} = \frac{\gamma}{P_{ti}}$, $\tilde{Y}_{ti} = \frac{A_{ti} y_{ti}}{P_{ti}}$ and $\alpha_{ti} = 2\gamma$. It is not hard to see then that the conditions of Lemma 12.1 are satisfied. In particular, for any $S \subset [K]$, $|S| > 1$, $\prod_{i \in S} A_{ti} = 0$, implying $\prod_{i \in S} \tilde{Y}_{ti} = 0$. Therefore

$$\mathbb{P}\left(\sum_{i=1}^K (\hat{L}_{ni} - L_{ni}) \geq \frac{\log(1/\delta')}{2\gamma}\right) \leq \delta'. \quad (12.8)$$

Similarly, for any fixed i ,

$$\mathbb{P}\left(\hat{L}_{ni} - L_{ni} \geq \frac{\log(1/\delta')}{2\gamma}\right) \leq \delta'. \quad (12.9)$$

To see this use the previous argument with $\alpha_{tj} = \mathbb{I}\{j = i\} 2\gamma$. The result follows by choosing $\delta' = \delta/(K+1)$ and the union bound. \square

LEMMA 12.3 $\tilde{L}_n - \hat{L}_n = \gamma \sum_{j=1}^K \hat{L}_{nj}$.

Proof Let $A_{ti} = \mathbb{I}\{A_t = i\}$ as before. Writing $Y_t = \sum_j A_{tj} y_{tj}$, we calculate

$$Y_t - \sum_{j=1}^K P_{tj} \hat{Y}_{tj} = \sum_{j=1}^K \left(1 - \frac{P_{tj}}{P_{tj} + \gamma}\right) A_{tj} y_{tj} = \gamma \sum_{j=1}^K \frac{A_{tj}}{P_{tj} + \gamma} y_{tj} = \gamma \sum_{j=1}^K \hat{Y}_{tj}.$$

Therefore $\tilde{L}_n - \hat{L}_n = \gamma \sum_{j=1}^K \hat{L}_{nj}$ as required. \square

Proof of Theorem 12.1 By Eq. (12.3) and Lemma 12.3 we have

$$\begin{aligned}\hat{R}_n &\leq \frac{\log(K)}{\eta} + (\tilde{L}_n - \hat{L}_n) + \max_{i \in [K]} (\hat{L}_{ni} - L_{ni}) + \frac{\eta}{2} \sum_{j=1}^K \hat{L}_{nj} \\ &= \frac{\log(K)}{\eta} + \max_{i \in [K]} (\hat{L}_{ni} - L_{ni}) + \left(\frac{\eta}{2} + \gamma\right) \sum_{j=1}^K \hat{L}_{nj}.\end{aligned}$$

Therefore by Lemma 12.2, with probability at least $1 - \delta$ it holds that

$$\begin{aligned}\hat{R}_n &\leq \frac{\log(K)}{\eta} + \frac{\log\left(\frac{K+1}{\delta}\right)}{2\gamma} + \left(\gamma + \frac{\eta}{2}\right) \left(\sum_{j=1}^K L_{nj} + \frac{\log\left(\frac{K+1}{\delta}\right)}{2\gamma} \right) \\ &\leq \frac{\log(K)}{\eta} + \left(\gamma + \frac{\eta}{2}\right) nK + \left(\gamma + \frac{\eta}{2} + 1\right) \log\left(\frac{K+1}{\delta}\right),\end{aligned}$$

where the second inequality follows since $L_{nj} \leq n$ for all j . The result follows by substituting the definitions of $\eta \in \{\eta_1, \eta_2\}$ and $\gamma = \eta/2$. \square

12.1.1 Proof of Lemma 12.1

We start with a technical inequality.

LEMMA 12.4 *For any $0 \leq x \leq 2\lambda$ it holds that $\exp\left(\frac{x}{1+\lambda}\right) \leq 1 + x$.*

Note that $1+x \leq \exp(x)$. What the lemma shows is that by slightly discounting the argument of the exponential function, in a bounded neighborhood of zero, $1+x$ can be an upper bound for the resulting function. Or, equivalently, slightly inflating the linear term in $1+x$, the linear lower bound becomes an upper bound.

Proof of Lemma 12.4 We rely on algebraic inequalities. The first is $\frac{2u}{1+u} \leq \log(1+2u)$ which holds for $u \geq 0$. The second states that $x \log(1+y) \leq \log(1+xy)$, which holds for any $x \in [0, 1]$ and $y > -1$. Thanks to these inequalities,

$$\frac{x}{1+\lambda} = \frac{x}{2\lambda} \frac{2\lambda}{1+\lambda} \leq \frac{x}{2\lambda} \log(1+2\lambda) \leq \log\left(1+2\lambda \frac{x}{2\lambda}\right) = \log(1+x).$$

And the proof is completed by exponentiating both sides. \square

Proof of Lemma 12.1 Fix $t \in [n]$ and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_t]$ denote the conditional expectation with respect to \mathcal{F}_t . By Lemma 12.4 and the assumption that $0 \leq \alpha_{ti} \tilde{Y}_{ti} \leq 2\lambda_{ti}$ we have

$$\exp\left(\frac{\alpha_{ti} \tilde{Y}_{ti}}{1+\lambda_{ti}}\right) \leq (1 + \alpha_{ti} \tilde{Y}_{ti}).$$

Taking the product of these inequalities over i ,

$$\begin{aligned} \mathbb{E}_{t-1} \left[\exp \left(\sum_{i=1}^K \frac{\alpha_{ti} \tilde{Y}_{ti}}{1 + \lambda_{ti}} \right) \right] &\leq \mathbb{E}_{t-1} \left[\prod_{i=1}^K (1 + \alpha_{ti} \tilde{Y}_{ti}) \right] \leq 1 + \mathbb{E}_{t-1} \left[\sum_{i=1}^K \alpha_{ti} \tilde{Y}_{ti} \right] \\ &= 1 + \sum_{i=1}^K \alpha_{ti} y_{ti} \leq \exp \left(\sum_{i=1}^K \alpha_{ti} y_{ti} \right), \end{aligned} \quad (12.10)$$

where the second inequality follows from the assumption that for $S \subset [K]$ with $|S| > 1$, $\mathbb{E}_{t-1} \prod_{i \in S} \tilde{Y}_{ti} \leq 0$, the third one follows from the assumption that $\mathbb{E}_{t-1} \tilde{Y}_{ti} = y_{ti}$, while the last one follows from $1 + x \leq \exp(x)$. Define

$$Z_t = \exp \left(\sum_i \alpha_{ti} \left(\frac{\tilde{Y}_{ti}}{1 + \lambda_{ti}} - y_{ti} \right) \right)$$

and let $M_t = Z_1 \dots Z_t$, $t \in [n]$ with $M_0 = 1$. By (12.10), $\mathbb{E}_{t-1}[Z_t] \leq 1$. Therefore

$$\mathbb{E}[M_t] = \mathbb{E}[\mathbb{E}_{t-1}[M_t]] = \mathbb{E}[M_{t-1} \mathbb{E}_{t-1}[Z_t]] \leq \mathbb{E}[M_{t-1}] \leq \dots \leq \mathbb{E}[M_0] = 1.$$

Setting $t = n$ and combining the above display with Markov's inequality leads to $\mathbb{P}(\log(M_n) \geq \log(1/\delta)) = \mathbb{P}(M_n \delta \geq 1) \leq \mathbb{E}[M_n] \delta \leq \delta$. \square

12.2 Notes

- 1 An upper bound on the expected regret of Exp3-IX can be obtained by integrating the tail.

$$R_n \leq \mathbb{E}[(\hat{R}_n)^+] = \int_0^\infty \mathbb{P}((\hat{R}_n)^+ \geq x) dx \leq \int_0^\infty \mathbb{P}(\hat{R}_n \geq x) dx,$$

where the first equality follows from Proposition 2.3. The result is completed using either high probability bound in Theorem 12.1 and by straightforward integration. We leave the details to the reader in Exercise 12.4.

- 2 The analysis presented here uses a fixed learning rate that depends on the horizon. Replacing η and γ with $\eta_t = \sqrt{\log(K)/(Kt)}$ and $\gamma_t = \eta_t/2$ leads to an anytime algorithm with about the same regret [Neu, 2015a].
- 3 There is another advantage of the modified importance-weighted estimators used by Exp3-IX, which leads to an improved regret in the special case that one of the arms has small losses. Specifically, it is possible to show that

$$R_n = O \left(\sqrt{K \min_{i \in [K]} L_{in} \log(K)} \right).$$

In the worst case L_{in} is linear in n and the usual bound is recovered. But if the optimal arm enjoys low cumulative regret, then the above can be a big improvement over the bounds given in Theorem 12.1. Bounds of this kind are called **first order bounds**. We refer the interested reader to the papers by Allenberg et al. [2006], Abernethy et al. [2012], Neu [2015b].

- 4 Another situation where one might hope to have a smaller regret is when the rewards/losses for each arm do not deviate too far from their averages. Define the **quadratic variation** by

$$Q_n = \sum_{t=1}^n \sqrt{\sum_{i=1}^K (x_{ti} - \mu_i)^2}, \quad \text{where } \mu_i = \frac{1}{n} \sum_{t=1}^n x_{ti}.$$

[Hazan and Kale \[2011\]](#) gave an algorithm for which $R_n = O(K^2\sqrt{Q_t})$, which can be better than the worst case bound of Exp3 or Exp3-IX when the quadratic variation is very small. The factor of K^2 is suboptimal and can be removed using a careful instantiation of the mirror descent algorithm [\[Bubeck et al., 2018\]](#). We do not cover this exact algorithm in this book, but the techniques based on mirror descent are presented in Chapter 28.

- 5 An alternative to the algorithm presented here is to mix the probability distribution computed using exponential weights with the uniform distribution, while biasing the estimates. This leads to the Exp3.P algorithm due to [Auer et al. \[2002b\]](#) who considered the case where δ is given and derived a bound that is similar to Eq. (12.6) of Theorem 12.1. With an appropriate modification of their proof it is possible to derive a weaker bound similar to Eq. (12.5) where the knowledge of δ is not needed by the algorithm. This has been explored by [Beygelzimer et al. \[2010\]](#) in the context of a related algorithm, which will be considered in Chapter 18. One advantage of this approach is that it generalizes to the case where the loss estimators are sometimes negative, a situation that can arise in more complicated settings. For technical details we advise the reader to work through Exercise 12.1.

12.3 Bibliographic remarks

The Exp3-IX algorithm is due to [Kocák et al. \[2014\]](#), who also introduced the biased loss estimators. The focus of that paper was to improve algorithms for more complex models with potentially large action-sets and side information, though their analysis can still be applied to the model studied in this chapter. The observation that this algorithm also leads to high probability bounds appeared in a followup paper by [Neu \[2015a\]](#). High probability bounds for adversarial bandits were first provided by [Auer et al. \[2002b\]](#) and explored in a more generic way by [Abernethy and Rakhlin \[2009\]](#). The idea to reduce the variance of importance-weighted estimators is not new and seems to have been applied in various forms [[Uchibe and Doya, 2004](#), [Wawrzynski and Pacut, 2007](#), [Ionides, 2008](#), [Bottou et al., 2013](#)]. All of these papers are based on truncating the estimators, which makes the resulting estimator less smooth. Surprisingly, the variance reduction technique used in this chapter seems to be recent [[Kocák et al., 2014](#)].

12.4 Exercises

12.1 In this exercise we ask you to analyze the Exp3.P algorithm, which as we mentioned in the notes is another way to obtain high probability bounds. The idea is to modify Exp3 by biasing the estimators and introducing some forced exploration. Let $\hat{Y}_{ti} = A_{ti}y_{ti}/P_{ti} - \eta/P_{ti}$ be a biased version of the loss-based importance-weighted estimator that was used in the previous chapter. Define $\hat{L}_{ti} = \sum_{s=1}^t \hat{Y}_{si}$ and consider the policy that samples $A_t \sim P_t$ where

$$P_{ti} = (1 - \gamma)\tilde{P}_{ti} + \frac{\gamma}{K} \quad \text{with} \quad \tilde{P}_{ti} = \frac{\exp(-\eta\hat{L}_{t-1,i})}{\sum_{j=1}^K \exp(-\eta\hat{L}_{t-1,j})}.$$

- (a) Let $\delta \in (0, 1)$ and $i \in [K]$. Show that with probability $1 - \delta$, the random regret \hat{R}_{ni} against i (cf. (12.2)) satisfies

$$\hat{R}_{ni} < n\gamma + (1 - \gamma) \sum_{t=1}^n \sum_{j=1}^K \tilde{P}_{tj}(\hat{Y}_{tj} - y_{ti}) + \sum_{t=1}^n \frac{\beta}{P_{tA_t}} + \sqrt{\frac{n \log(1/\delta)}{2}}.$$

- (b) Show that

$$\sum_{t=1}^n \sum_{j=1}^K \tilde{P}_{tj}(\hat{Y}_{tj} - y_{ti}) = \sum_{t=1}^n \sum_{j=1}^K \tilde{P}_{tj}(\hat{Y}_{tj} - \hat{Y}_{ti}) + \sum_{t=1}^n (\hat{Y}_{ti} - y_{ti}).$$

- (c) Show that

$$\sum_{t=1}^n \sum_{j=1}^K \tilde{P}_{tj}(\hat{Y}_{tj} - \hat{Y}_{ti}) \leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^n \sum_{j=1}^K \tilde{P}_{tj}\hat{Y}_{tj}^2.$$

- (d) Show that

$$\sum_{t=1}^n \sum_{j=1}^K \tilde{P}_{tj}\hat{Y}_{tj}^2 \leq \frac{nK^2\beta^2}{\gamma} + \sum_{t=1}^n \frac{1}{P_{tA_t}}.$$

- (e) Apply the result of Exercise 5.17 to show that for any $\delta \in (0, 1)$, the following hold:

$$\begin{aligned} \mathbb{P} \left(\sum_{t=1}^n \frac{1}{P_{tA_t}} \geq 2nK + \frac{K}{\gamma} \log \left(\frac{1}{\delta} \right) \right) &\leq \delta. \\ \mathbb{P} \left(\sum_{t=1}^n \hat{Y}_{ti} - y_{ti} \geq \frac{1}{\beta} \log \left(\frac{1}{\delta} \right) \right) &\leq \delta. \end{aligned}$$

- (f) Combining the previous steps, show that there exists a universal constant $C > 0$ such that for any $\delta \in (0, 1)$, for an appropriate choice of η, γ and β , with probability at least $1 - \delta$ it holds that that the random regret \hat{R}_n of Exp3.P satisfies

$$\hat{R}_n \leq C \sqrt{nK \log(K/\delta)}$$

- (g) In which step did you use the modified estimators?
- (h) Show a bound where the algorithm parameters η, γ, β can only depend on n, K , but not on δ .
- (i) Compare the bounds with the analogous bounds for Exp3-IX in Theorem 12.1.

12.2 This exercise is concerned with a generalization of the core idea underlying Exp3.P of the previous exercise in that rather than giving explicit expressions for the biased loss estimates, we focus on the key properties of these that makes Exp3.P “tick”. To reduce clutter we assume for the remainder that t ranges in $[n]$ and $a \in [K]$. Let $(\Omega, \mathcal{F}, \mathcal{G} \doteq (\mathcal{G}_t)_{t=0}^n, \mathbb{P})$ be a filtered probability space. Let (Z_t) , (\hat{Z}_t) , (\tilde{Z}_t) , (β_t) be sequences of random elements in \mathbb{R}^K , where $\tilde{Z}_t = \hat{Z}_t - \beta_t$ and $(Z_t), (\beta_t)$ are \mathcal{G} -predictable, whereas (\hat{Z}_t) and therefore also (\tilde{Z}_t) are \mathcal{G} -adapted (think of \hat{Z}_t as the estimate of Z_t that uses randomization, and β_t is the bias as in the previous exercise). Given positive constant η define the probability vector $P_t \in \mathcal{P}_{K-1}$ by

$$P_{ta} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{sa}\right)}{\sum_{b=1}^K \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{sb}\right)}.$$

Let $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot | \mathcal{G}_{t-1}]$. Assume the following hold for all $a \in [K]$:

- (a) $\eta |\hat{Z}_{ta}| \leq 1$, (b) $\eta \beta_{ta} \leq 1$,
 (c) $\eta \mathbb{E}_{t-1}[\hat{Z}_{ta}^2] \leq \beta_{ta}$ almost surely, (d) $\mathbb{E}_{t-1}[\hat{Z}_{ta}] = Z_{ta}$ almost surely.

Let $A^* = \operatorname{argmin}_{a \in [K]} \sum_{t=1}^n Z_{ta}$ and $R_n = \sum_{t=1}^n \sum_{a=1}^K P_{ta}(Z_{ta} - Z_{tA^*})$.

- (a) Show that

$$\begin{aligned} & \sum_{t=1}^n \sum_{a=1}^K P_{ta}(Z_{ta} - Z_{tA^*}) \\ &= \underbrace{\sum_{t=1}^n \sum_{a=1}^K P_{ta}(\tilde{Z}_{ta} - \tilde{Z}_{tA^*})}_{(A)} + \underbrace{\sum_{t=1}^n \sum_{a=1}^K P_{ta}(Z_{ta} - \tilde{Z}_{ta})}_{(B)} + \underbrace{\sum_{t=1}^n (\tilde{Z}_{tA^*} - Z_{tA^*})}_{(C)}. \end{aligned}$$

- (b) Show that

$$(A) \leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^n \sum_{a=1}^K P_{ta} \hat{Z}_{ta}^2 + 3 \sum_{t=1}^n \sum_{a=1}^K P_{ta} \beta_{ta}.$$

- (c) Show that with probability at least $1 - \delta$,

$$(B) \leq 2 \sum_{t=1}^n \sum_{a=1}^K P_{ta} \beta_{ta} + \frac{\log(1/\delta)}{\eta}.$$

(d) Show that with probability at least $1 - K\delta$,

$$(C) \leq \frac{\log(1/\delta)}{\eta}.$$

(e) Conclude that for any $\delta \leq 1/(K+1)$, with probability at least $1 - (K+1)\delta$,

$$R_n \leq \frac{3\log(1/\delta)}{\eta} + \eta \sum_{t=1}^n \sum_{a=1}^K P_{ta} \hat{Z}_{ta}^2 + 5 \sum_{t=1}^n \sum_{a=1}^K P_{ta} \beta_{ta}.$$

 This is a long and challenging exercise. You may find it helpful to use the result in Exercise 5.17. The solution is also available.

12.3 Consider the Bernoulli bandit with $K = 5$ arms and $n = 10^4$ with means $\mu_1 = 1/2$ and $\mu_i = 1/2 - \Delta$ for $i > 1$. Plot the regret of Exp3 and Exp3-IX for $\Delta \in [0, 1/2]$. You should get something similar to the graph in Fig. 12.1. Does the result surprise you? Repeat the experiment in Part (k) of Exercise 11.5 with Exp3-IX and convince yourself that this algorithm is more robust than Exp3.

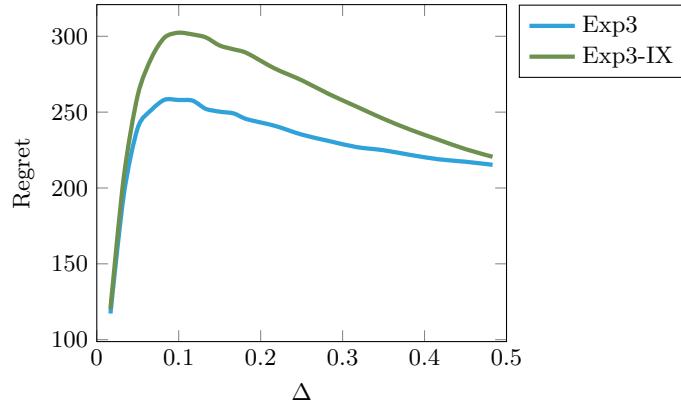


Figure 12.1 Comparison between Exp3 and Exp3-IX on Bernoulli bandit

12.4 In this exercise you will complete the steps explained in Note 1 to prove a bound on the expected regret of Exp3-IX.

(a) Find a choice of η and universal constant $C > 0$ such that

$$R_n \leq C\sqrt{Kn \log(K)}.$$

(b) What happens as η grows? Write a bound on the expected regret of Exp3-IX in terms of η and K and n .

Part IV

Lower Bounds for Bandits with Finitely Many Arms

Until now we have indulged ourselves by presenting algorithms and upper bounds on their regret. As satisfying as this is, the real truth of a problem is usually to be found in the lower bounds. There are several reasons for this:

- 1 An upper bound does not tell you much about what you could be missing out on. The only way to demonstrate that your algorithm really is (close to) optimal is to prove a lower bound showing that no algorithm can do better.
- 2 The second reason that lower bounds are often more informative in the sense that it usually turns out to be easier to get the lower bound right than the upper bound. History shows a list of algorithms with steadily improving guarantees until eventually someone hits upon the idea for which the upper bound matches some known lower bound.
- 3 Finally, thinking about lower bounds forces you to understand what is hard about the problem. This is so useful that the best place to start when attacking a new problem is usually to try and prove lower bounds. Too often we have not heeded our own advice and started trying to design an algorithm, only to discover later that had we tackled the lower bound first, then the right algorithm would have fallen in our laps with almost no effort at all.

So what is the form of a typical lower bound? In the chapters that follow we will see roughly two flavours. The first is the worst case lower bound, which corresponds to a claim of the form

“For any policy you give me, I will give you an instance of a bandit problem ν on which the regret is at least L ”.

Results of this kind have an adversarial flavour, which makes them suitable for understanding the robustness of a policy. The second type is a lower bound on the regret of an algorithm for specific instances. These bounds have a different form that usually reads like the following:

“If you give me a *reasonable* policy, then its regret on any instance ν is at least $L(\nu)$ ”.

The statement only holds for some policies – the ‘reasonable’ ones, whatever that means. But the guarantee is also more refined because bound controls the regret for these policies on every instance by a function that depends on this instance. This kind of bound will allow us to show that the instance-dependent bounds for stochastic bandits of $O(\sum_{i:\Delta_i > 0} \Delta_i + \log(n)/\Delta_i)$ are not improvable. The inclusion of the word ‘reasonable’ is unfortunately necessary. For every bandit instance ν there is a policy that just chooses the optimal action in ν . Such policies are not reasonable because they have linear regret for bandits with a different optimal arm. In the chapters that follow we will see various ways to define ‘reasonable’ in a way that is simultaneously rigorous and, well, reasonable.

The contents of this part is roughly as follows. First we introduce the definition of worst case regret and discuss the line of attack for proving lower bounds (Chapter 13). The next chapter takes us on a brief excursion into information theory where we explain the necessary mathematical tools (Chapter 14). Readers

familiar with information theory could skim this chapter. The final three chapters are devoted to applying information theory to prove lower bounds on the regret for both stochastic and adversarial bandits.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

13 Lower Bounds: Basic Ideas

We start the block on lower bounds by considering stochastic bandits. Let \mathcal{E} be a set of stochastic bandits and π be a policy. The **worst case regret** of policy π on environment class \mathcal{E} is

$$R_n(\pi, \mathcal{E}) = \sup_{\nu \in \mathcal{E}^K} R_n(\pi, \nu).$$

Let Π be the set of all policies. The **minimax regret** is

$$R_n^*(\mathcal{E}) = \inf_{\pi \in \Pi} R_n(\pi, \mathcal{E}) = \inf_{\pi \in \Pi} \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu).$$

A policy is called **minimax optimal** for \mathcal{E} if $R_n(\pi, \mathcal{E}^K) = R_n^*(\mathcal{E})$. The value $R_n^*(\mathcal{E})$ is of interest by itself. A small value of $R_n^*(\mathcal{E})$ indicates that the underlying bandit problem is less challenging in the worst-case sense. A core activity in bandit theory is to understand what makes $R_n^*(\mathcal{E})$ large or small, often focusing on its behavior as a function of the number of rounds n .



Minimax optimality is not a property of a policy alone. It is a property of a policy together with a set of environments and a horizon.

Finding a minimax policy is generally too computationally expensive to be practical. For this reason we almost always settle for a policy that is nearly minimax optimal. One of the main results of this part is a proof of the following theorem, which together with Theorem 9.1 shows that Algorithm 6 from Chapter 9 is minimax optimal up to constant factors for 1-subgaussian bandits with suboptimality gaps in $[0, 1]$.

THEOREM 13.1 *Let \mathcal{E}^K be the set of K -armed Gaussian bandits with unit variance and means $\mu \in [0, 1]^K$. Then there exists a constant $c > 0$ such that for all $K > 1$ and $n \geq K$ it holds that*

$$R_n^*(\mathcal{E}^K) \geq c\sqrt{(K-1)n}.$$

We will prove this theorem in Chapter 15, but first we give an informal justification. Let X_1, X_2, \dots, X_n be an observed sequence of independent Gaussian random variables with unknown mean μ and known variance 1. Assume you are told that μ takes on one of two values: $\mu = 0$ or $\mu = \Delta$ for some known $\Delta > 0$.

Your task is to guess the value of μ . Let $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$ be the sample mean, which is Gaussian with mean μ and variance $1/n$. While it is not immediately obvious how easy this task is, intuitively we expect the optimal decision is to predict that $\mu = 0$ if $\hat{\mu}$ is closer to 0 than to Δ , and otherwise to predict $\mu = \Delta$. For large n we expect our prediction will probably be correct. Supposing that $\mu = 0$ (the other case is symmetric), then the prediction will be wrong only if $\hat{\mu} \geq \Delta/2$. Using the fact that $\hat{\mu}$ is Gaussian with mean $\mu = 0$ and variance $1/n$, combined with known bounds on the Gaussian tail probabilities (see [Abramowitz and Stegun, 1964](#)) leads to

$$\begin{aligned} \frac{1}{\sqrt{n\Delta^2 + \sqrt{n\Delta^2 + 4}}} \sqrt{\frac{2}{\pi}} \exp\left(-\frac{n\Delta^2}{8}\right) &\leq \mathbb{P}\left(\hat{\mu} \geq \frac{\Delta}{2}\right) \\ &\leq \frac{1}{\sqrt{n\Delta^2 + \sqrt{n\Delta^2 + 8/\pi}}} \sqrt{\frac{2}{\pi}} \exp\left(-\frac{n\Delta^2}{8}\right). \end{aligned}$$

The upper and lower bounds are incredibly close, differing only in the constant in the square root of the denominator. One might believe that the decision procedure could be improved, but the symmetry of the problem makes this seem improbable. The formula exhibits the expected behaviour, which is that once n is large relative to $8/\Delta^2$, then the probability that this procedure fails drops exponentially with further increases in n . But the lower bound also shows that if n is small relative to $8/\Delta^2$, then the procedure fails with constant probability.

The problem described is called hypothesis testing and the ideas underlying the argument above are core to many impossibility result in statistics. The next task is to reduce our bandit problem to hypothesis testing. The high level idea is to select two bandit problem instances in such a way that the following two conditions hold simultaneously:

- 1 *Competition*: A sequence of actions that is good for one bandit is not good for the other.
- 2 *Similarity*: The instances are ‘close’ enough that the policy interacting with either of the two instances cannot statistically identify the true bandit with reasonable statistical accuracy.

The two requirements are clearly conflicting. The first makes us want to choose instances with means $\mu, \mu' \in [0, 1]^K$ that are far from each other, while the second requirement makes us want to choose them to be close to each other. The lower bound will follow by optimizing this tradeoff.

Let us start to make things concrete by choosing bandits $\nu = (P_i)_{i=1}^K$ and $\nu' = (P'_i)_{i=1}^K$ where $P_i = \mathcal{N}(\mu_i, 1)$ and $P'_i = \mathcal{N}(\mu'_i, 1)$ are Gaussian and $\mu, \mu' \in [0, 1]^K$. In order to prove a lower bound it suffices to show that for every strategy π there exists a choice of μ and μ' such that

$$\max \{R_n(\pi, \nu), R_n(\pi, \nu')\} \geq c\sqrt{Kn},$$

where $c > 0$ is a universal constant. Let $\Delta > 0$ be a constant to be tuned

subsequently and choose $\mu = (\Delta, 0, 0, \dots, 0)$, which means that the first arm is optimal in instance ν and

$$R_n(\pi, \nu) = (n - \mathbb{E}[T_1(n)])\Delta, \quad (13.1)$$

where the expectation is taken with respect to the induced measure on the sequence of outcomes when π interacts with ν . Now we need to choose μ' to satisfy the two requirements above. Since we want ν and ν' to be hard to distinguish and yet have different optimal actions, we should make μ' as close to μ except in a coordinate where π expects to explore the least. To this end, let

$$i = \operatorname{argmin}_{i>1} \mathbb{E}[T_i(n)]$$

be the suboptimal arm in ν that π expects to play least often. By the pigeonhole principle and the fact that $\sum_i \mathbb{E}[T_i(n)] = n$, it must hold that

$$\mathbb{E}[T_i(n)] \leq \frac{n}{K-1}.$$

Then define $\mu' \in \mathbb{R}^K$ by

$$\mu'_j = \begin{cases} \mu_j & \text{if } j \neq i \\ 2\Delta & \text{otherwise.} \end{cases}$$

The regret in this bandit is

$$R_n(\pi, \nu') = \Delta \mathbb{E}'[T_1(n)] + \sum_{j \notin \{1, i\}} 2\Delta \mathbb{E}'[T_j(n)] \geq \Delta \mathbb{E}'[T_1(n)], \quad (13.2)$$

where $\mathbb{E}'[\cdot]$ is the expectation operator on the sequence of outcomes when π interacts with ν' . So now we have the following situation: The strategy π interacts with either ν or ν' and when interacting with ν it expects to play arm i at most $n/(K-1)$ times. But the two instances only differ when playing arm i . The time has come to tune Δ . Because the strategy expects to play arm i only about $n/(K-1)$ times, taking inspiration from the previous discussion on distinguishing samples from Gaussian distributions with different means, we will choose

$$\Delta = \sqrt{\frac{1}{\mathbb{E}[T_i(n)]}} \geq \sqrt{\frac{K-1}{n}}.$$

If we are prepared to ignore the fact that $T_i(n)$ is a random variable and take for granted the claims in the first part of the chapter, then with this choice of Δ the strategy cannot distinguish between instances ν and ν' and in particular we expect that $\mathbb{E}[T_1(n)] \approx \mathbb{E}'[T_1(n)]$. If $\mathbb{E}[T_1(n)] \leq n/2$, then by Eq. (13.1) we have

$$R_n(\pi, \nu) = \Delta \mathbb{E}[T_1(n)] \geq \frac{n}{2} \sqrt{\frac{K-1}{n}} = \frac{1}{2} \sqrt{n(K-1)}.$$

On the other hand, if $\mathbb{E}[T_i(n)] \geq n/2$, then

$$R_n(\pi, \nu') \geq \Delta \mathbb{E}'[T_i(n)] \approx \Delta \mathbb{E}[T_i(n)] \geq \frac{1}{2} \sqrt{n(K-1)},$$

which completes our heuristic argument that there exists a universal constant $c > 0$ such that

$$R_n^*(\mathcal{E}^K) \geq c\sqrt{nK}.$$

We have been sloppy in many places: The claims in the first part of the chapter have not been proven yet and $T_i(n)$ is a random variable. Before we can present the rigorous argument we need a chapter to introduce some ideas from information theory. Readers already familiar with these concepts can skip to Chapter 15 for the proof of Theorem 13.1.

13.1 Notes

- 1 The worst-case regret has a game-theoretic interpretation. Imagine a game between a protagonist and an antagonist that works as follows: For $K > 1$ and $n \geq K$ the protagonist proposes a bandit policy π . The antagonist looks at the policy and chooses a bandit ν from the class of environments considered. The utility for the antagonist is the expected regret and for the protagonist it is the negation of the expected regret, which makes this a zero-sum game. Both players aim to maximize their payoffs. The game is completely described by n and \mathcal{E} . One characteristic value in a game is its minimax value. As described above, this is a sequential game (the protagonist moves first, then the antagonist). The minimax value of this game from the perspective of the antagonist is exactly $R_n^*(\mathcal{E})$, while for the protagonist is $\sup_{\pi} \inf_{\nu} (-R_n(\pi, \nu)) = -R_n^*(\mathcal{E})$.
- 2 We mentioned that finding the minimax optimal policy is usually computationally infeasible. In fact it is not clear we should even try. In classical statistics it often turns out that minimizing the worst case leads to a flat risk profile. In the language of bandits this would mean that the regret is the same for every bandit (where possible). What we usually want in practice is to have low regret against ‘easy’ bandits and larger regret against ‘hard’ bandits. The analysis in Part II suggests that easy bandits are those where the suboptimality gaps are large or very small. There is evidence to suggest that the exact minimax optimal strategy may not exploit these easy instances, so in practice one might prefer to find a policy that is nearly minimax optimal and has much smaller regret on easy bandits. We will tackle questions of this nature in Chapter 16.
- 3 The regret on a class of bandits \mathcal{E} is a multi-objective criteria. Some policies will be good for some instances and bad on others, and there are clear trade-offs. One way to analyze the performance in a multi-objective setting is called **Pareto optimality**. A policy is Pareto optimal if there does not exist another policy that is a strict improvement. More precisely, if there does not exist a π' such that $R_n(\pi', \nu) \leq R_n(\pi, \nu)$ for all $\nu \in \mathcal{E}$ and $R_n(\pi', \nu) < R_n(\pi, \nu)$ for at least one instance $\nu \in \mathcal{E}$.

4 When we say a policy is minimax optimal up to constant factors for finite-armed 1-subgaussian bandits with suboptimality gaps in $[0, 1]$ we mean there exists a $C > 0$ such that

$$\frac{R_n(\pi, \mathcal{E}^K)}{R_n^*(\mathcal{E}^K)} \leq C \text{ for all } K \text{ and } n,$$

where \mathcal{E}^K is the set of K -armed 1-subgaussian bandits with suboptimality gaps in $[0, 1]$. We often say a policy is minimax optimal up to logarithmic factors, by which we mean that

$$\frac{R_n(\pi, \mathcal{E}^K)}{R_n^*(\mathcal{E}^K)} \leq C(n, K) \text{ for all } K \text{ and } n,$$

where $C(n, K)$ is logarithmic in n and K . We hope the reader will forgive us for not always specifying in the text exactly what is meant and promise that statements of theorems will always be precise.

13.2 Exercises

13.1 Let $\mathbb{P}_\mu = \mathcal{N}(\mu, 1)$ be the Gaussian measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ with mean $\mu \in \{0, \Delta\}$ and unit variance. Let $X : \mathbb{R} \rightarrow \mathbb{R}$ be the identity random variable ($X(\omega) = \omega$). For decision rule $d : \mathbb{R} \rightarrow \{0, \Delta\}$ define risk

$$R(d) = \max_{\mu \in \{0, \Delta\}} \mathbb{P}_\mu(d(X) \neq \mu),$$

Prove that $R(d)$ is minimized by $d(x) = \operatorname{argmin}_{\tilde{\mu} \in \{0, \mu\}} |x - \tilde{\mu}|$.

13.2 Let $K > 1$ and $\mathcal{E} = \mathcal{E}_{\mathcal{N}}^K(1)$ be the set of Gaussian bandits with unit variance. Find a Pareto optimal policy for this class.



Think about simple policies (not necessarily good ones) and use the definition.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

14 Foundations of Information Theory (†)

To make the arguments in the previous chapter rigorous and generalizable to other settings we need some classic tools from information theory and statistics. In particular, we will need the concept of **relative entropy**, also known as the **Kullback-Leibler divergence** named for Solomon Kullback and Richard Leibler (KL divergence, for short).

The relative entropy has several interpretations. The one we will focus on here comes from the situation encountered by Alice, who wants to communicate with Bob. She wants to tell Bob the outcome of a sequence of independent random variables sampled from known distribution Q . Alice and Bob agree to communicate using a code that is fixed in advance in such a way that the expected message length is minimized. Then the **entropy** of Q is the expected length of the optimal code. The relative entropy between distributions P and Q is the price in terms of expected message length that Alice and Bob have to pay if they believe the random variables are sampled from Q , when in fact they are sampled from P .

Let X be a random variable that takes finitely many values, which without loss of generality we will assume is $X \in [N]$. We abbreviate $p_i = \mathbb{P}(X = i)$. Let us first discuss how to define the amount of information that observing X conveys. One way to start is to define information as the amount of communication needed if we want to tell a friend about the value we observed. We'll assume that Alice observes the value of X and wants to tell Bob what value she observed using a **binary code** that they agree upon in advance. For example, if $N = 4$, then they might agree on the following code: $1 \rightarrow 00, 2 \rightarrow 01, 3 \rightarrow 10, 4 \rightarrow 11$. Then if Alice observes a 3, she sends Bob a message containing 10. For our purposes, a code is a function $c : [N] \rightarrow \{0, 1\}^*$ where $\{0, 1\}^*$ is the set of finite sequences of zeros and ones.

Of course we demand that c is injective so that no two numbers (or **symbols**) have the same code. We also require that c is **prefix free**, which means that no code should be the prefix of any other. This is justified by supposing that Alice would like to tell Bob about multiple samples. Then Bob needs to know where the message for one symbol starts and ends, and he would like to do this with no back-tracking.

The easiest choice is to use $\lceil \log_2(N) \rceil$ bits no matter the value of X . This simple code is sometimes effective, but is not entirely satisfactory if X is far from uniform.

To understand why, suppose that N is extremely large and $\mathbb{P}(X = 1) = 0.99$ and the remaining probability mass is uniform over $[N] - \{1\}$. Then it seems preferable to have a short code for 1 and slightly longer codes for the alternatives. With this in mind, a natural objective is to find a code that minimizes the expected code length. That is

$$c = \operatorname{argmin}_c \sum_{i=1}^N p_i \ell(c(i)), \quad (14.1)$$

where the argmin is taken over valid codes and $\ell(\cdot)$ is a function that returns the length of a code. The optimization problem in (14.1) can be solved in using Huffman coding and the value lies within the following range

$$H_2(P) \leq \min_c \sum_{i=1}^N p_i \ell(c(i)) \leq H_2(P) + 1,$$

where $H_2(P)$ is defined by

$$H_2(P) = \sum_{i \in [N]: p_i > 0} p_i \log_2 \left(\frac{1}{p_i} \right).$$

Notice that if P is uniform, then $p_i = 1/N$ and the naive idea of using a code of uniform length is recovered, but for non-uniform distributions the code adapts to assign shorter codes to symbols with high probability. What is not apparent from the expression above is that the code length for symbol i when using Huffman coding is never longer than $\log(1/p_i) + 1$. It is also worth pointing out that the sum is only over outcomes that occur with non-zero probability, which is motivated by observing that $\lim_{x \rightarrow 0} x \log(1/x) = 0$ or by thinking of the entropy as an expectation of the log-probability with respect to P and expectations should not change when the value of the random variable is perturbed on a measure zero set.

It turns out that $H_2(P)$ is not just an approximation on the expected length of the Huffman code, but is itself a fundamental quantity. We will not go into this in detail, but imagine that Alice wants to send a long string of symbols to Bob. She could use a Huffman code to send Bob each symbol one at a time, but this introduces ‘rounding errors’ that accumulate as the message length grows. Instead they can agree on a procedure, which Bob can still interpret sequentially without backtracking, and for which the expected average code-length averaged over the whole message tends towards $H_2(P)$ as the length of the message grows. Furthermore, the celebrated source coding theorem says that you cannot do better than this! For reference, a procedure for achieving this called **arithmetic coding**, but for our purposes we do not actually want to code messages, but rather to understand the meaning of information.

Before moving on, we will replace the base 2 logarithm with its natural

counterpart and define the entropy of a random variable X by

$$H(P) = \sum_{i \in [N]: p_i > 0} p_i \log\left(\frac{1}{p_i}\right). \quad (14.2)$$

This is nothing more than a scaling of the H_2 that is ultimately mathematically more convenient. Measuring information using base 2 logarithms has a unit of **bits** and for the natural logarithm it is called **nats**.

We hope you agree that $H(P)$ measures the (expected) information content of observing random variables sampled from P , at least in the long run. We now move towards defining the relative entropy.

14.1 The relative entropy

Suppose that Alice and Bob agree to use a code that is optimal when X is sampled from distribution Q . Unbeknownst to them, however, let us suppose that actually X is sampled from distribution P . The relative entropy between P and Q measures how much longer the messages are expected to be using the optimal code for Q than what would be obtained using the optimal code for P . Letting $p_i = P(X = i)$ and $q_i = Q(X = i)$ and working out the math leads to the definition of the relative entropy as

$$D(P, Q) = \sum_{i \in [N]: p_i > 0} p_i \log\left(\frac{1}{q_i}\right) - \sum_{i \in [N]: p_i > 0} p_i \log\left(\frac{1}{p_i}\right) = \sum_{i \in [N]: p_i > 0} p_i \log\left(\frac{p_i}{q_i}\right) \quad (14.3)$$

If this quantity is large, then we expect to be able to tell that $P \neq Q$ with fewer independent observations sharing P than if this quantity was smaller. For example, if there exists an i with $p_i > 0$ and $q_i = 0$, then the first time we see symbol i , we can tell with certainty that the symbol was not sampled from Q . Looking at the definition of the relative entropy shows that in this case, $D(P, Q) = \infty$.

Still poking around the definition, what happens when $q_i = 0$ and $p_i = 0$? This means that the symbol i is superfluous and the value of $D(P, Q)$ should not be impacted by introducing superfluous symbols. And again, it does not by the definition of the expectations. We also see that the sufficient and necessary condition for $D(P, Q) < \infty$ is that for each i such that $q_i = 0$, we also have that $p_i = 0$. The condition we discovered is also expressed as saying that P is absolutely continuous with respect to Q , which is also written as $P \ll Q$. Note that absolute continuity only implies a finite relative entropy when X takes on finitely many values. If instead $X \in \{2, 3, 4, \dots\}$ and $\mathbb{P}(X = i) \propto 1/(i \log^2(i))$, then $H(X) = \infty$.

More generally, for two measures P, Q on a common measurable space (Ω, \mathcal{F}) , we say that P is **absolutely continuous** with respect to Q (and write $P \ll Q$) if for any $A \in \mathcal{F}$, $Q(A) = 0$ implies that $P(A) = 0$ (intuitively, \ll is like \leq except

that it only constrains the values when the right-hand side is also zero). This brings us back to defining relative entropy between two arbitrary probability distributions P, Q defined over a common probability space. The difficulty we face is that if $X \sim P$ takes on uncountably infinitely many values then we cannot really use the ideas that use communication because no matter what coding we use, we would need infinitely many symbols to describe some values of X . How can then be the entropy of X be defined at all? This seems to be a truly fundamental difficulty! Luckily, this impasse gets resolved automatically if we only consider relative entropy. While we cannot communicate X , for any *finite* ‘discretization’ of the the possible values that X can take on, the discretized values can be communicated finitely and all our definitions will work. Formally, if X takes values in the measurable space $(\mathcal{X}, \mathcal{G})$, with \mathcal{X} possibly having uncountably many elements, a discretization to $[N]$ levels would be specified using some function $f : \mathcal{X} \rightarrow [N]$ that is $\mathcal{G}/2^{[N]}$ -measurable map. Then, the entropy of P relative Q , $D(P, Q)$ can be defined as

$$D(P, Q) = \sup_f D(P_f, Q_f),$$

where P_f is the distribution of $Y = f(X)$ when $X \sim P$ and Q_f is the distribution of $Y = f(X)$ when $X \sim Q$ and the supremum is for all $N \in \mathbb{N}^+$ and all maps f as defined above. In words, we take all possible discretizations f (with no limit on the ‘fineness’ of the discretization) and define $D(P, Q)$ as the excess information when expecting to see $f(X)$ with $X \sim Q$ while reality is $X \sim P$. If this is finite, then we expect this to be a reasonable definition. As we shall see it soon, it is indeed a reasonable definition.

THEOREM 14.1 *Let (Ω, \mathcal{F}) be a measurable space and let P and Q be measures on this space. Then,*

$$D(P, Q) = \begin{cases} \int \log \left(\frac{dP}{dQ}(\omega) \right) dP(\omega), & \text{if } P \ll Q; \\ \infty, & \text{otherwise.} \end{cases}$$

Note that by our earlier remark, this reduces to (14.3) for discrete measures. If λ is a common dominating σ -finite measure for P and Q (that is, $P \ll \lambda$ and $Q \ll \lambda$ both hold) then letting $p = \frac{dP}{d\lambda}$ and $q = \frac{dQ}{d\lambda}$, if also $P \ll Q$, the chain rule gives $\frac{dP}{dQ} \frac{dQ}{d\lambda} = \frac{dP}{d\lambda}$, which lets us write

$$D(P, Q) = \int p \log \left(\frac{p}{q} \right) d\lambda,$$

which is perhaps the best known expression for relative entropy and is also often used as a definition. Note that for probability measures, a common dominating σ -finite measure can always be bound. For example, $\lambda = P + Q$ always dominates both P and Q .

Relative entropy is a kind of ‘distance’ measure between distributions P and Q . In particular, if $P = Q$, then $D(P, Q) = 0$ and otherwise $D(P, Q) > 0$. Strictly

speaking, the relative entropy is not a distance because it satisfies neither the triangle inequality nor is it symmetric. Nevertheless, it serves the same purpose.

The relative entropy between many standard distributions is often quite easy to compute. For example, the relative entropy between two Gaussians with means $\mu_1, \mu_2 \in \mathbb{R}$ and common variance σ^2 is

$$D(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}.$$

The dependence on the difference in means and the variance is consistent with our intuition. If μ_1 is close to μ_2 , then the ‘difference’ between the distributions should be small, but if the variance is very small, then there is little overlap and the difference is large. The relative entropy between two Bernoulli distributions with means $p, q \in [0, 1]$

$$D(\mathcal{B}(p), \mathcal{B}(q)) = p \log\left(\frac{p}{q}\right) + (1-p) \log\left(\frac{1-p}{1-q}\right),$$

where $0 \log(\cdot) = 0$.

We are nearing the end of our whirlwind tour of relative entropy. It remains to state the key lemma, sometimes called the **high probability Pinsker** inequality, that connects the relative entropy to the hardness of hypothesis testing.

THEOREM 14.2 *Let P and Q be probability measures on the same measurable space (Ω, \mathcal{F}) and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp(-D(P, Q)), \quad (14.4)$$

where $A^c = \Omega \setminus A$ is the complement of A .

The proof may be found at the end of the chapter, but first some interpretation and a simple application. Suppose that $D(P, Q)$ is small, then P is ‘close’ to Q in some sense. Since P is a probability measure we have $P(A) + P(A^c) = 1$. If Q is close to P , then we might expect $P(A) + Q(A^c)$ should be large. The purpose of the theorem is to quantify just how large. Note that if P is not absolutely continuous with respect to Q then $D(P, Q) = \infty$ and the result is vacuous. Also note that the result is symmetric. We could replace $D(P, Q)$ with $D(Q, P)$, which sometimes leads to a stronger result because the relative entropy is not symmetric.

Returning to the hypothesis testing problem described in the previous chapter. Let X be normally distributed with unknown mean $\mu \in \{0, \Delta\}$ and variance $\sigma^2 > 0$. We want to bound the quality of a rule for deciding what is the real mean from a single observation. The decision rule is characterized by a measurable set $A \subseteq \mathbb{R}$ on which the predictor guesses $\mu = \Delta$ (it predicts $\mu = 0$ on the complement of A). Let $P = \mathcal{N}(0, \sigma^2)$ and $Q = \mathcal{N}(\Delta, \sigma^2)$. Then the probability of an error under P is $P(A)$ and the probability of error under Q is $Q(A^c)$. The reader surely knows what to do next. By Theorem 14.2 we have

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp(-D(P, Q)) = \frac{1}{2} \exp\left(-\frac{\Delta^2}{2\sigma^2}\right).$$

If we assume that the signal to noise ratio is small, $\Delta^2/\sigma^2 \leq 1$, then

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp\left(-\frac{1}{2}\right) \geq \frac{3}{10},$$

which implies $\max\{P(A), Q(A^c)\} \geq 3/20$. This means that no matter how we chose our decision rule, we simply do not have enough data to make a decision for which the probability of error on either P or Q is smaller than $3/20$.

$$P(A) + Q(A^c) \geq 1 - \sqrt{\frac{1}{2} D(P, Q)}$$

Proof of Theorem 14.2 For reals a, b we abbreviate $\max\{a, b\} = a \vee b$ and $\min\{a, b\} = a \wedge b$. The result is trivial if $D(P, Q) = \infty$. On the other hand, by Theorem 14.1, $D(P, Q) < \infty$ implies that $P \ll Q$. Let $\nu = P + Q$. Then $P, Q \ll \nu$, which by Theorem 2.5 ensures the existence of the Radon-Nikodym derivatives $p = \frac{dP}{d\nu}$ and $q = \frac{dQ}{d\nu}$. The chain rule gives

$$\frac{dP}{dQ} \frac{dQ}{d\nu} = \frac{dP}{d\nu} \quad \text{and} \quad \frac{dP}{dQ} = \frac{\frac{dP}{d\nu}}{\frac{dQ}{d\nu}}.$$

Therefore

$$D(P, Q) = \int p \log\left(\frac{p}{q}\right) d\nu.$$

For brevity, when writing integrals with respect to ν , in this proof, we will drop $d\nu$. Thus, we will write, for example $\int p \log(p/q)$ for the above integral. Instead of (14.4), we prove the stronger result that

$$\int p \wedge q \geq \frac{1}{2} \exp(-D(P, Q)). \quad (14.5)$$

This indeed is sufficient since $\int p \wedge q = \int_A p \wedge q + \int_{A^c} p \wedge q \leq \int_A p + \int_{A^c} q = P(A) + Q(A^c)$. We start with an inequality attributed to French mathematician Lucien Le Cam, which lower bounds the left-hand side of Eq. (14.5). The inequality states that

$$\int p \wedge q \geq \frac{1}{2} \left(\int \sqrt{pq} \right)^2. \quad (14.6)$$

Starting from the right-hand side above using $pq = (p \wedge q)(p \vee q)$ and Cauchy-Schwartz we get

$$\left(\int \sqrt{pq} \right)^2 = \left(\int \sqrt{(p \wedge q)(p \vee q)} \right)^2 \leq \left(\int p \wedge q \right) \left(\int p \vee q \right).$$

Now, using $p \wedge q + p \vee q = p + q$, the proof is finished by substituting $\int p \vee q = 2 - \int p \wedge q \leq 2$ and dividing both sides by two.

Thus, it remains to lower bound the right-hand side of (14.6). For this, we use

Jensen's inequality. First, we write $(\cdot)^2$ as $\exp(2\log(\cdot))$ and then move the log inside the integral:

$$\begin{aligned} \left(\int \sqrt{pq}\right)^2 &= \exp\left(2\log\int \sqrt{pq}\right) = \exp\left(2\log\int p\sqrt{\frac{q}{p}}\right) \\ &\geq \exp\left(2\int p\frac{1}{2}\log\left(\frac{q}{p}\right)\right) = \exp\left(-\int_{pq>0} p\log\left(\frac{p}{q}\right)\right) \\ &= \exp\left(-\int p\log\left(\frac{p}{q}\right)\right) = \exp(-D(P, Q)). \end{aligned}$$

In the fourth and the last step we used that since $P \ll Q$, $q = 0$ implies $p = 0$ and so $p > 0$, which implies $q > 0$, and eventually $pq > 0$. The result is completed by chaining the inequalities. \square

14.2 Notes

- 1 Theorem 14.1 connects our definition of relative entropies to densities (the ‘classic definition’). It can be found in Section 5.2 of the book by Gray [2011].
- 2 How tight is Theorem 14.2? We remarked already that $D(P, Q) = 0$ if and only if $P = Q$. But in this case Theorem 14.2 only gives

$$1 = P(A) + Q(A^c) \geq \frac{1}{2} \exp(-D(P, Q)) = \frac{1}{2},$$

which does not seem so strong. From where does the weakness arise? The answer is in Eq. (14.6), which can be refined by

$$\left(\int \sqrt{pq}\right)^2 \leq \left(\int p \wedge q\right) \left(\int p \vee q\right) = \left(\int p \wedge q\right) \left(2 - \int p \wedge q\right)$$

By solving the quadratic inequality we have

$$\begin{aligned} P(A) + Q(A^c) &\geq \int p \wedge q \geq 1 - \sqrt{1 - \left(\int \sqrt{pq}\right)^2} \\ &\geq 1 - \sqrt{1 - \exp(-D(P, Q))}, \end{aligned} \tag{14.7}$$

which gives a modest improvement on Theorem 14.2 that becomes more pronounced when $D(P, Q)$ is close to zero as demonstrated by Fig. 14.1. This stronger bound might be useful for fractionally improving constant factors in lower bounds, but we do not know of any application for which it is really crucial and the more complicated form makes it cumbersome to use. Part of the reason for this is that the situation where $D(P, Q)$ is small is better dealt with using an even stronger inequality (for just that case). See the next note!

- 3 Another inequality from information theory is **Pinsker’s inequality**, which states for measures P and Q on the same probability space (Ω, \mathcal{F}) that

$$\delta(P, Q) = \sup_{A \in \mathcal{F}} P(A) - Q(A) \leq \sqrt{\frac{1}{2} D(P, Q)}. \tag{14.8}$$

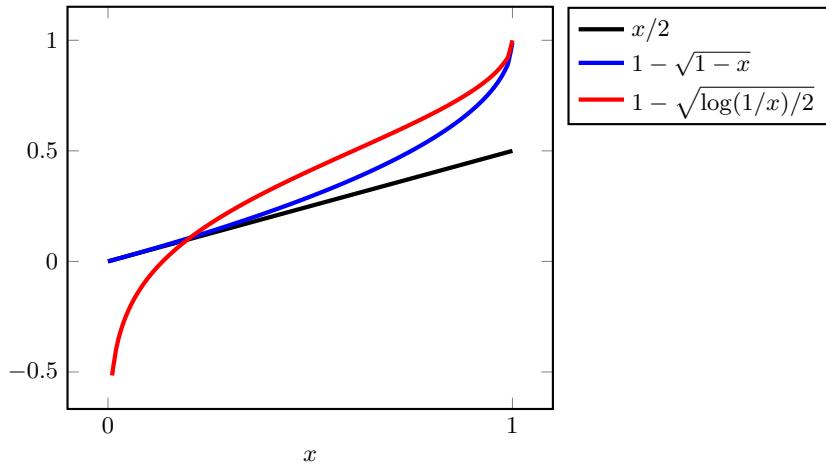


Figure 14.1 Tightening the inequality of Le Cam

As an aside, the quantity on the left hand side is called the **total variation distance** between P and Q , which actually is a distance on the space of probability measures (on the same probability space, of course). From this we can derive for any measurable $A \in \mathcal{F}$ that

$$P(A) + Q(A^c) \geq 1 - \sqrt{\frac{1}{2} D(P, Q)} = 1 - \sqrt{\frac{1}{2} \log \left(\frac{1}{\exp(-D(P, Q))} \right)}.$$

Examining Fig. 14.1 shows that this is an improvement on Eq. (14.7) when $D(P, Q)$ is small.

- 4 We saw the total variation distance in Eq. (14.8). There are two other ‘distances’ that are occasionally useful. These are the **Hellinger distance** and the χ^2 -**distance**, which using the notation in the proof of Theorem 14.2 are defined by

$$h(P, Q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2} = \sqrt{2 \left(1 - \int \sqrt{pq} \right)} \quad (14.9)$$

$$\chi^2(P, Q) = \int \frac{(p - q)^2}{q} = \int \frac{p^2}{q} - 1. \quad (14.10)$$

Notice that $h(P, Q)$ is bounded and exists for all probability measures P and Q , while a necessary condition for the χ^2 -distance to exist is that $P \ll Q$. Like the total variation distance, the Hellinger distance is actually a distance (it is symmetric and satisfies triangle inequality), but the χ^2 -‘distance’ is not. It is possible to show (see Chapter 2 of the book by [Tsybakov \[2008\]](#)) that

$$\delta(P, Q)^2 \leq h(P, Q)^2 \leq D(P, Q) \leq \chi^2(P, Q). \quad (14.11)$$

Each of the inequalities are tight for some choices of P and Q , but the examples

do not chain together as evidenced by Pinsker's inequality, which shows that $\delta(P, Q)^2 \leq D(P, Q)/2$ (which is also tight for some P and Q).

- 5 Let $P = (p_i)_i$ be a distribution on $[N]$. Another interpretation of the entropy $H(P)$ is as a measure of the amount of uncertainty in P . But what do we mean by uncertainty? One approach to define uncertainty is to think of how much one should be surprised to see a particular value of X (sampled from P). If x is deterministic, then there is no surprise at all and so the uncertainty measure should be zero. And indeed, $H(P) = 0$ when P is a Dirac measure. On the other hand, if X is uniformly distributed, then we should be equally surprised by any value, which provides some support defining the amount of 'surprise' when observing $X = i$ by $\log(1/p_i)$. Then entropy is the 'expected surprise'. Long story short, it turns out that reasonable definitions of uncertainty actually give rise to the definition of H in Eq. (14.2).
- 6 The entropy for distribution P was defined as $H(P)$ in Eq. (14.2). If X is a random variable, then $H(X)$ is defined to be the entropy of the law of X . This is a convenient notation because it allows one to write $H(f(X))$ and $H(XY)$ and similar expressions.

14.3 Bibliographic remarks

There are many references for information theory. Most well known (and comprehensive) is the book by [Cover and Thomas \[2012\]](#). Another famous book is the elementary and enjoyable introduction by [MacKay \[2003\]](#). The approach we have taken for defining and understanding the relative entropy is inspired by an excellent shorter book by [Gray \[2011\]](#).

14.4 Exercises

14.1 Let (Ω, \mathcal{F}) be a measurable space and let $P, Q : \mathcal{F} \rightarrow [0, 1]$ be probability measures. Let $a < b$ and $X : \Omega \rightarrow [a, b]$ be a \mathcal{F} -measurable random variable. Prove that

$$\left| \int_{\Omega} X(\omega) dP(\omega) - \int_{\Omega} X(\omega) dQ(\omega) \right| \leq (b-a)\delta(P, Q).$$

14.2 Prove that each of the inequalities in Eq. (14.11) is tight.

14.3 Let Ω be a countable set and $p : \Omega \rightarrow [0, 1]$ be a distribution on Ω so that $\sum_{\omega \in \Omega} p(\omega) = 1$. Let P be the measure associated with p , which means that $P(A) = \sum_{\omega \in A} p(\omega)$. The **counting measure** μ is the measure on $(\Omega, 2^{\Omega})$ given by $\mu(A) = |A|$ if A is finite and $\mu(A) = \infty$ otherwise.

- (a) Show that P is absolutely continuous with respect to μ .
- (b) Show that the Radon-Nykodim $dP/d\mu$ exists and that $dP/d\mu(\omega) = p(\omega)$.

14.4 For each $i \in \{1, 2\}$ let $\mu_i \in \mathbb{R}$, $\sigma_i^2 > 0$ and $P_i = \mathcal{N}(\mu_i, \sigma_i^2)$. Show that

$$\mathcal{D}(P_1, P_2) = \frac{1}{2} \left(\log \left(\frac{\sigma_2^2}{\sigma_1^2} \right) + \frac{\sigma_1^2}{\sigma_2^2} - 1 \right) + \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2}.$$

14.5 Let $(\mathbb{R}, \mathfrak{L})$ be a measurable space with \mathfrak{L} the Lebesgue σ -algebra and let λ be the Lebesgue measure. Find:

- a probability measure $(\mathbb{R}, \mathfrak{L})$ that is *not* absolutely continuous with respect to λ .
- a probability measure P on $(\mathbb{R}, \mathfrak{L})$ that *is* absolutely continuous to λ with $\mathcal{D}(P, Q) = \infty$ where $Q = \mathcal{N}(0, 1)$ is the standard Gaussian measure.

14.6 Let P and Q be measures on (Ω, \mathcal{F}) and let \mathcal{G} be a sub- σ -algebra of \mathcal{F} and $P_{\mathcal{G}}$ and $Q_{\mathcal{G}}$ be the restrictions of P and Q to (Ω, \mathcal{G}) . Show that $\mathcal{D}(P_{\mathcal{G}}, Q_{\mathcal{G}}) \leq \mathcal{D}(P, Q)$.

15 Minimax Lower Bounds

After the short excursion into information theory, let us return to the world of K -armed stochastic bandits. In what follows we fix the horizon $n > 0$ and the number of actions $K > 1$. This chapter has two components. The first is an exact calculation of the relative entropy between measures in the canonical bandit model for a fixed policy and different bandits. In the second component we prove a minimax lower bound that formalizes the intuitive arguments given in Chapter 13.

15.1 Relative entropy between bandits

The following result will be used repeatedly. Some generalizations are provided in the notes and exercises.

LEMMA 15.1 (Divergence decomposition) *Let $\nu = (P_1, \dots, P_K)$ be the reward distributions associated with one K -armed bandit, and let $\nu' = (P'_1, \dots, P'_K)$ be the reward distributions associated with the another K -armed bandit. Fix some policy π and let $\mathbb{P}_\nu = \mathbb{P}_{\nu\pi}$ and $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu'\pi}$ be the measures on the canonical bandit model (Section 4.4) induced by the interconnection of π and ν (respectively, π and ν'). Then*

$$D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{i=1}^K \mathbb{E}_\nu[T_i(n)] D(P_i, P'_i). \quad (15.1)$$

Proof Assume that $D(P_i, P'_i) < \infty$ for all $i \in [K]$. From this it follows that $P_i \ll P'_i$. Define $\lambda = \sum_{i=1}^K P_i + P'_i$, which is the measure defined by $\lambda(A) = \sum_{i=1}^K (P_i(A) + P'_i(A))$ for any measurable set A . Recalling that ρ is the counting measure on $[K]$, Theorem 14.1 shows that

$$D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \mathbb{E}_\nu \left[\log \left(\frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}} \right) \right].$$

The Nikodym derivative of \mathbb{P}_ν with respect to the product measure $(\rho \times \lambda)^n$ is given in Eq. (4.5) as

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

The density of $\mathbb{P}_{\nu'}$ is identical except that p_{a_t} is replaced by p'_{a_t} . Then

$$\log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}(a_1, x_1, \dots, a_n, x_n) = \sum_{t=1}^n \log \frac{p_{a_t}(x_t)}{p'_{a_t}(x_t)},$$

where we used the chain rule for Radon-Nikodym derivatives and the fact that the terms involving the policy cancel. Taking expectations of both sides:

$$\mathbb{E}_\nu \left[\log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}(A_t, X_t) \right] = \sum_{t=1}^n \mathbb{E}_\nu \left[\log \frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)} \right],$$

and

$$\mathbb{E}_\nu \left[\log \frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)} \right] = \mathbb{E}_\nu \left[\mathbb{E}_\nu \left[\log \frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)} \mid A_t \right] \right] = \mathbb{E}_\nu [\text{D}(P_{A_t}, P'_{A_t})],$$

where in the second equality we used that under $\mathbb{P}_\nu(\cdot | A_t)$ the distribution of X_t is $dP_{A_t} = p_{A_t} d\lambda$. Plugging back into the previous display,

$$\begin{aligned} \mathbb{E}_\nu \left[\log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}(A_1, X_1, \dots, A_n, X_n) \right] &= \sum_{t=1}^n \mathbb{E}_\nu \left[\log \frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)} \right] \\ &= \sum_{t=1}^n \mathbb{E}_\nu [\text{D}(P_{A_t}, P'_{A_t})] = \sum_{i=1}^K \mathbb{E}_\nu \left[\sum_{t=1}^n \mathbb{I}\{A_t = i\} \text{D}(P_{A_t}, P'_{A_t}) \right] \\ &= \sum_{i=1}^K \mathbb{E}_\nu [T_i(n)] \text{D}(P_i, P'_i). \end{aligned}$$

When the right-hand side of (15.1) is infinite, by our previous calculation, it is not hard to see that the left-hand side will also be infinite. \square

15.2 Minimax lower bounds

Recall that $\mathcal{E}_N^K(1)$ is the class of Gaussian bandits with unit variance, which can be parameterized by their mean vector $\mu \in \mathbb{R}^K$. Given $\mu \in \mathbb{R}^K$ let ν_μ be the Gaussian bandit for which the i th arm has reward distribution $\mathcal{N}(\mu_i, 1)$.

THEOREM 15.1 *Let $K > 1$ and $n \geq K - 1$. Then for any policy π there exists a mean vector $\mu \in [0, 1]^K$ such that*

$$R_n(\pi, \nu_\mu) \geq \frac{1}{27} \sqrt{(K-1)n}.$$

Since $\nu_\mu \in \mathcal{E}_K$, it follows that the minimax regret for \mathcal{E}_K is lower bounded by the right-hand side of the above display as soon as $n \geq K - 1$:

$$R_n^*(\mathcal{E}_K) \geq \frac{1}{27} \sqrt{(K-1)n}.$$

The idea of the proof is illustrated in Fig. 15.1.

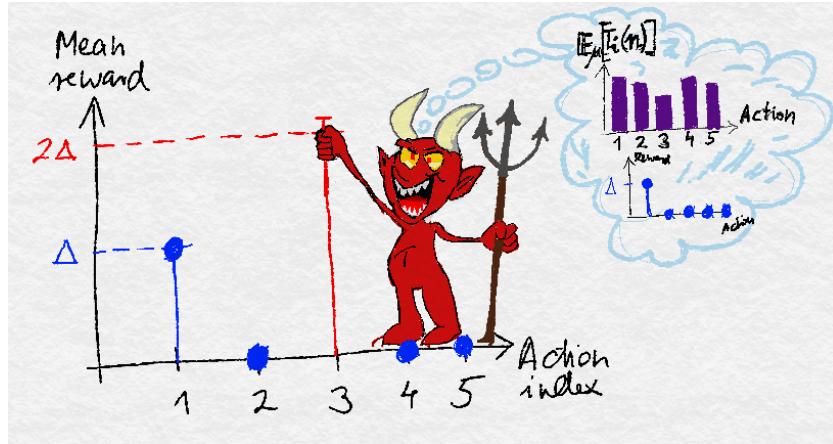


Figure 15.1 The idea of the minimax lower bound. Given a policy and one environment, the evil antagonist picks another environment so that the policy will suffer a large regret in at least one environment

Proof Fix a policy π . Let $\Delta \in [0, 1/2]$ be some constant to be chosen later. As suggested in Chapter 13 we start with a Gaussian bandit with unit variance and mean vector $\mu = (\Delta, 0, 0, \dots, 0)$. This environment and π gives rise to the distribution $\mathbb{P}_{\nu_\mu, \pi}$ on the canonical bandit model $(\mathcal{H}_n, \mathcal{F}_n)$. For brevity we will use \mathbb{P}_μ in place of $\mathbb{P}_{\nu_\mu, \pi}$ and expectations under \mathbb{P}_μ will be denoted by \mathbb{E}_μ . To choose the second environment, let

$$i = \operatorname{argmin}_{j>1} \mathbb{E}_\mu[T_j(n)].$$

Since $\sum_{j=1}^K \mathbb{E}_\mu[T_j(n)] = n$, it holds that $\mathbb{E}_\mu[T_i(n)] \leq n/(K-1)$. The second bandit is also Gaussian with unit variance and means

$$\mu' = (\Delta, 0, 0, \dots, 0, 2\Delta, 0, \dots, 0),$$

where specifically $\mu'_i = 2\Delta$. Therefore $\mu_j = \mu'_j$ except at index i and the optimal optimal arm in ν_μ is the first arm and in i in $\nu_{\mu'}$. We abbreviate $\mathbb{P}_{\mu'} = \mathbb{P}_{\nu_{\mu'}, \pi}$. Lemma 4.2 and a simple calculation leads to

$$R_n(\pi, \nu_\mu) \geq \mathbb{P}_\mu(T_1(n) \leq n/2) \frac{n\Delta}{2} \quad \text{and} \quad R_n(\pi, \nu_{\mu'}) > \mathbb{P}_{\mu'}(T_1(n) > n/2) \frac{n\Delta}{2}.$$

Then applying the high probability Pinsker inequality from the previous chapter (Theorem 14.2),

$$\begin{aligned} R_n(\pi, \nu_\mu) + R_n(\pi, \nu_{\mu'}) &> \frac{n\Delta}{2} (\mathbb{P}_\mu(T_1(n) \leq n/2) + \mathbb{P}_{\mu'}(T_1(n) > n/2)) \\ &\geq \frac{n\Delta}{4} \exp(-D(\mathbb{P}_\mu, \mathbb{P}_{\mu'})). \end{aligned}$$

It remains to upper bound $D(\mathbb{P}_\mu, \mathbb{P}_{\mu'})$. For this, we use Lemma 15.1 and the

definitions of μ and μ' to get

$$D(\mathbb{P}_\mu, \mathbb{P}_{\mu'}) = \mathbb{E}_\mu[T_i(n)] D(\mathcal{N}(0, 1), \mathcal{N}(2\Delta, 1)) = \mathbb{E}_\mu[T_i(n)] \frac{(2\Delta)^2}{2} \leq \frac{2n\Delta^2}{K-1}.$$

Plugging this into the previous display, we find that

$$R_n(\pi, \nu_\mu) + R_n(\pi, \nu_{\mu'}) \geq \frac{n\Delta}{4} \exp\left(-\frac{2n\Delta^2}{K-1}\right).$$

The result is completed by choosing $\Delta = \sqrt{(K-1)/4n} \leq 1/2$, where the inequality follows from the assumptions in the theorem statement. The final steps are lower bounding $\exp(-1/2)$ and using $2 \max(a, b) \geq a + b$. \square

We encourage readers to go through the alternative proof outlined in Exercise 15.1, which takes a slightly different path.

15.3 Notes

- 1 We used the Gaussian noise model because the KL divergences are so easily calculated in this case, but all that we actually used was that $D(P_i, P'_i) = O((\mu_i - \mu'_i)^2)$ when the gap between the means $\Delta = \mu_i - \mu'_i$ is small. While this is certainly not true for *all* distributions, it very often is. Why is that? Let $\{P_\mu : \mu \in \mathbb{R}\}$ be some parametric family of distributions on Ω and assume that distribution P_μ has mean μ . Assuming the densities are twice differentiable and that everything is sufficiently nice that integrals and derivatives can be exchanged (as is almost always the case), we can use a Taylor expansion about μ to show that

$$\begin{aligned} D(P_\mu, P_{\mu+\Delta}) &\approx \frac{\partial}{\partial \Delta} D(P_\mu, P_{\mu+\Delta}) \Big|_{\Delta=0} \Delta + \frac{1}{2} \frac{\partial^2}{\partial \Delta^2} D(P_\mu, P_{\mu+\Delta}) \Big|_{\Delta=0} \Delta^2 \\ &= \frac{\partial}{\partial \Delta} \int_{\Omega} \log\left(\frac{dP_\mu}{dP_{\mu+\Delta}}\right) dP_\mu \Big|_{\Delta=0} \Delta + \frac{1}{2} I(\mu) \Delta^2 \\ &= - \int_{\Omega} \frac{\partial}{\partial \Delta} \log\left(\frac{dP_{\mu+\Delta}}{dP_\mu}\right) \Big|_{\Delta=0} dP_\mu \Delta + \frac{1}{2} I(\mu) \Delta^2 \\ &= - \int_{\Omega} \frac{\partial}{\partial \Delta} \frac{dP_{\mu+\Delta}}{dP_\mu} \Big|_{\Delta=0} dP_\mu \Delta + \frac{1}{2} I(\mu) \Delta^2 \\ &= - \frac{\partial}{\partial \Delta} \int_{\Omega} \frac{dP_{\mu+\Delta}}{dP_\mu} dP_\mu \Big|_{\Delta=0} \Delta + \frac{1}{2} I(\mu) \Delta^2 \\ &= - \frac{\partial}{\partial \Delta} \int_{\Omega} dP_{\mu+\Delta} \Big|_{\Delta=0} \Delta + \frac{1}{2} I(\mu) \Delta^2 \\ &= \frac{1}{2} I(\mu) \Delta^2, \end{aligned}$$

where $I(\mu)$, introduced in the second line, is called the **Fisher information** of the family $(P_\mu)_\mu$ at μ . Note that if λ is a common dominating measure for

$(P_{\mu+\Delta})$ for Δ small, $dP_{\mu+\Delta} = p_{\mu+\Delta} d\lambda$ and we can write

$$I(\mu) = - \int \frac{\partial^2}{\partial \Delta^2} \log p_{\mu+\Delta} \Big|_{\Delta=0} p_\mu d\lambda,$$

which is the form that is usually given in elementary texts. The upshot of all this is that $D(P_\mu, P_{\mu+\Delta})$ for Δ small is indeed quadratic in Δ , with the scaling provided by $I(\mu)$, and as a result the worst-case regret is always $O(\sqrt{nK})$, provided the class of distributions considered is sufficiently rich and not too bizarre.

- 2 We have now shown a lower bound that is $\Omega(\sqrt{nK})$, while many of the upper bounds were $O(\log(n))$. There is no contradiction because the logarithmic bounds depended on the inverse suboptimality gaps, which may be vary large.
- 3 Our lower bound was only proven for $n \geq K - 1$. In Exercise 15.2 we ask you to show that when $n < K - 1$ there exists a bandit such that

$$R_n \geq \frac{n(2K - n - 1)}{2K} > \frac{n}{2}.$$

15.4 Bibliographic remarks

The first work on lower bounds that we know of was the remarkably precise minimax analysis of two-armed Gaussian bandits by [Vogel \[1960\]](#). The high probability Pinsker inequality (Theorem 14.2) was first used for bandits by [Bubeck et al. \[2013b\]](#), but the theorem has other applications. As far as we can tell, the earliest proof is due to [Bretagnolle and Huber \[1979\]](#), but we also recommend the book by [Tsybakov \[2008\]](#). The proof of Theorem 15.1 uses the same ideas as [Gerchinovitz and Lattimore \[2016\]](#), while the alternative proof in Exercise 15.1 is essentially due to [Auer et al. \[1995\]](#), who analyzed the more difficult case where the rewards are Bernoulli (see Exercise 15.3).

15.5 Exercises

15.1 There is another way to prove Theorem 15.1. Let $c > 0$ and $\Delta = 2c\sqrt{K/n}$ and for each $i \in \{0, 1, \dots, K\}$ let $\mu^{(i)} \in \mathbb{R}^K$ satisfy $\mu_k^{(i)} = \mathbb{I}\{i = k\} \Delta$. Further abbreviate the notation in the proof of Theorem 15.1 by letting $\mathbb{E}_i[\cdot] = \mathbb{E}_{\mu^{(i)}}[\cdot]$.

- (a) Use Pinsker's inequality (Eq. 14.8) and Lemma 15.1 and the result of Exercise 14.1 to show

$$\mathbb{E}_i[T_i(n)] \leq \mathbb{E}_0[T_i(n)] + n\sqrt{\frac{1}{4}\Delta^2\mathbb{E}_0[T_i(n)]} = \mathbb{E}_0[T_i(n)] + c\sqrt{nK\mathbb{E}_0[T_i(n)]}.$$

- (b) Using the previous part, Jensen's inequality and the identity $\sum_{i=1}^K \mathbb{E}_0[T_i(n)] = n$, show that

$$\sum_{i=1}^K \mathbb{E}_i[T_i(n)] \leq n + c \sum_{i=1}^K \sqrt{nK \mathbb{E}_0[T_i(n)]} \leq n + cKn.$$

- (c) Let $R_i = R_n(\pi, G_{\mu^{(i)}})$. Find a choice of $c > 0$ for which

$$\begin{aligned} \sum_{i=1}^K R_i &= \Delta \sum_{i=1}^K (n - \mathbb{E}_i[T_i(n)]) \geq \Delta_i (nK - n - cKn) \\ &= 2c\sqrt{\frac{K}{n}} (nK - n - cKn) \geq \frac{nK}{8} \sqrt{\frac{K}{n}} \end{aligned}$$

- (d) Conclude there exists an $i \in [K]$ such that

$$R_i \geq \frac{1}{8} \sqrt{Kn}.$$

15.2 Let $K > 1$ and $n < K$. Prove that for any policy π there exists a Gaussian bandit with unit variance and means $\mu \in [0, 1]^K$ such that $R_n(\pi, \nu_\mu) \geq n(2K - n - 1)/(2K) > n/2$.

15.3 Recall from Table 4.1 that $\mathcal{E}_{\mathcal{B}^K}$ is the set of K -armed Bernoulli bandits. Show that there exists a universal constant $c > 0$ such that for any $2 \leq K \leq n$ it holds that:

$$R_n^*(\mathcal{E}_{\mathcal{B}^K}) = \inf_{\pi} \sup_{\nu \in \mathcal{E}_{\mathcal{B}^K}} R_n(\pi, \nu) \geq c\sqrt{nK}.$$



Use the fact that KL divergence is upper bounded by the χ^2 -distance (14.11).

15.4 In Chapter 9 we proved that if π is the MOSS policy and $\nu \in \mathcal{E}_{\text{SG}}^K(1)$, then

$$R_n(\pi, \nu) \leq C \left(\sqrt{Kn} + \sum_{i: \Delta_i > 0} \Delta_i \right),$$

where $C > 0$ is a universal constant. Prove that the dependence on the sum cannot be eliminated.



You will have to use that $T_i(t)$ is an integer for all t .

15.5 Let ETC_{nm} be the Explore-Then-Commit policy with inputs n and m respectively (Algorithm 1). Prove that for all m there exists a $\mu \in [0, 1]^K$ such that

$$R_n(\text{ETC}_{nm}, \nu_\mu) \geq c \min \left\{ n, n^{2/3} K^{1/3} \right\},$$

where $c > 0$ is a universal constant.

15.6 Consider the setting of Lemma 15.1 and let X be a random variable and $\mathbb{P}_{\nu X}$ and $\mathbb{P}_{\nu' X}$ be the distributions of X induced by \mathbb{P}_ν and $\mathbb{P}_{\nu'}$ respectively. Let $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$ and τ be a \mathcal{F}_t -measurable stopping time. Show that if X is \mathcal{F}_τ -measurable, then

$$D(\mathbb{P}_{\nu X}, \mathbb{P}_{\nu' X}) \leq \sum_{i=1}^K \mathbb{E}_\nu[T_i(\tau)] D(P_i, P'_i).$$

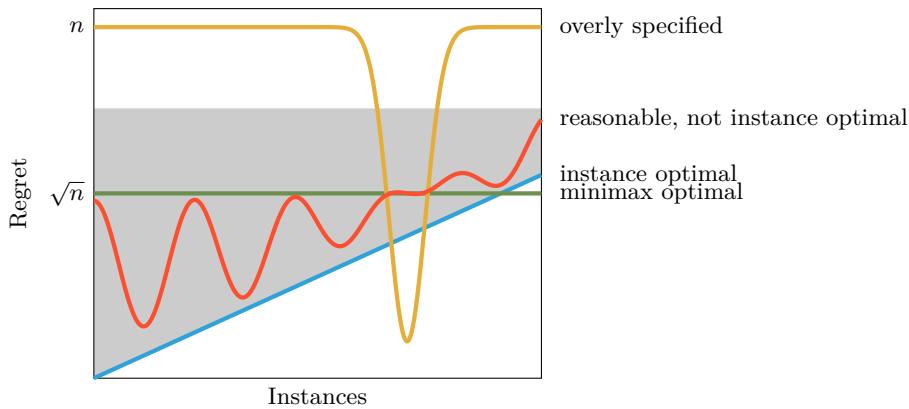
This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

16 Instance Dependent Lower Bounds

In the last chapter we proved a lower bound on the minimax regret for subgaussian bandits with suboptimality gaps in $[0, 1]$. Such bounds serve as a useful measure of the robustness of a policy, but are often excessively conservative. This chapter is devoted to understanding **instance-dependence** lower bounds, which try to capture the optimal performance of a policy on a specific bandit instance.

Because the regret is a multi-objective criteria, an algorithm designer might try and design algorithms that perform well on one kind of instance or another. An extreme example is the policy that chooses $A_t = 1$ for all t , which suffers zero regret when the first arm is optimal and linear regret otherwise. This is a harsh tradeoff with the price for reducing the regret from logarithmic to zero on just a few instances being linear regret on the remainder. Surprisingly, this is the nature of the game in bandits. One can assign a measure of difficulty to each instance such that policies performing overly well relative to this measure on some instances pay a steep price on others. The situation is illustrated in the figure below.



On the x -axis the instances are ordered according to the measure of difficulty and the y -axis shows the regret (on some scale). In the previous chapter we proved that no policy can be entirely below the horizontal ‘minimax optimal’ line. The results in this chapter show that if the regret of a policy is below the ‘instance optimal’ line at any point, then it must have regret above the shaded region for other instances. For example, the ‘overly specified’ policy. In finite-time the

situation is a little messy, but if one pushes these ideas to the limit, then for many classes of bandits one can define a precise notion of instance-dependent optimality.

16.1 Asymptotic bounds

We need to define exactly what is meant by a reasonable policy. If one is only concerned with asymptotics, then a rather conservative definition suffices.

DEFINITION 16.1 A policy π is called **consistent** over a class of bandits \mathcal{E} if for all $\nu \in \mathcal{E}$ and $p > 0$ it holds that

$$\lim_{n \rightarrow \infty} \frac{R_n(\pi, \nu)}{n^p} = 0. \quad (16.1)$$

The class of consistent policies over \mathcal{E} is denoted by $\Pi_{\text{cons}}(\mathcal{E})$.

Theorem 7.1 shows that UCB is consistent over $\mathcal{E}_{\text{SG}}^K(1)$. The strategy that always chooses the first action is not consistent on any class \mathcal{E} unless $K = 1$ or \mathcal{E} is so restrictive that the first arm is optimal action for every $\nu \in \mathcal{E}$.



Consistency is an asymptotic notion. A policy could be consistent and yet play $A_t = 1$ for all $t \leq 10^{100}$. For this reason an assumption on consistency is insufficient to derive nonasymptotic lower bounds. Later we introduce a finite-time version of consistency that allows us to prove finite-time instance-dependent lower bounds.

A class \mathcal{E} of stochastic bandits is **unstructured** if $\mathcal{E} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_K$ with $\mathcal{C}_1, \dots, \mathcal{C}_K$ sets of distributions. The main theorem of this chapter is a generic lower bound that applies to any unstructured class of stochastic bandits. After the proof we will see some applications to specific classes. Let \mathcal{C} be a set of distributions with finite means and let $\mu : \mathcal{C} \rightarrow \mathbb{R}$ be the function that maps $P \in \mathcal{C}$ to its mean. Let $\alpha \in \mathbb{R}$ and $P \in \mathcal{C}$ have $P(\mu) < \alpha$ and define

$$d_{\mathcal{C}}(P, \alpha) = \inf_{P' \in \mathcal{C}} \{D(P, P') : \mu(P') > \alpha\}.$$

Recall that $P_i(\nu)$ is the distribution of rewards for the i th arm of bandit ν and $\mu_i(\nu)$ is its mean and $\mu^*(\nu) = \max_i \mu_i(\nu)$ and $\Delta_i(\nu) = \mu^*(\nu) - \mu_i(\nu)$.

THEOREM 16.1 *Let $\mathcal{E} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_K$ and $\pi \in \Pi_{\text{cons}}(\mathcal{E})$ be a consistent policy over \mathcal{E} . Then for all $\nu \in \mathcal{E}$ it holds that*

$$\liminf_{n \rightarrow \infty} \frac{R_n}{\log(n)} \geq c^*(\nu, \mathcal{E}) = \sum_{i: \Delta_i(\nu) > 0} \frac{\Delta_i(\nu)}{d_{\mathcal{C}_i}(P_i(\nu), \mu^*(\nu))}. \quad (16.2)$$

Proof Abbreviate $P_i = P_i(\nu)$ and $\mu_i = \mu_i(\nu)$ and $\Delta_i = \Delta_i(\nu)$ and $\mu^* = \mu^*(\nu)$ and $d_i = d_{\mathcal{C}_i}(P_i, \mu^*)$. The result will follow from Lemma 4.2 and by showing that for any suboptimal arm i it holds that

$$\liminf_{n \rightarrow \infty} \frac{\mathbb{E}_{\nu\pi}[T_i(n)]}{\log(n)} \geq \frac{1}{d_i}.$$

Fix a suboptimal arm i and let $\varepsilon > 0$ be arbitrary and $\nu' \in \mathcal{E}$ be a bandit with $P_j(\nu') = P_j$ for $j \neq i$ and $P_i(\nu')$ be such that $D(P_i, P'_i) \leq d_i + \varepsilon$ and $\mu(P'_i) > \mu^*$, which exists by the definition of d_i . Then by Lemma 15.1 we have $D(\mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu',\pi}) \leq \mathbb{E}_{\nu\pi}[T_i(n)](d_i + \varepsilon)$ and by Theorem 14.2 for any event A

$$\mathbb{P}_{\nu\pi}(A) + \mathbb{P}_{\nu',\pi}(A) \geq \frac{1}{2} \exp(-D(\mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu',\pi})) \geq \frac{1}{2} \exp(-\mathbb{E}_{\nu\pi}[T_i(n)](d_i + \varepsilon)).$$

Now choose $A = \{T_i(n) > n/2\}$ and let $R_n = R_n(\pi, \nu)$ and $R'_n = R_n(\pi, \nu')$. Then

$$\begin{aligned} R_n + R'_n &\geq \frac{n}{2} (\mathbb{P}_{\nu\pi}(A)\Delta_i + \mathbb{P}_{\nu',\pi}(A^c)(\mu'_i - \mu^*)) \\ &\geq \frac{n}{2} \min\{\Delta_i, \mu'_i - \mu^*\} (\mathbb{P}_{\nu\pi}(A) + \mathbb{P}_{\nu',\pi}(A^c)) \\ &\geq \frac{n}{2} \min\{\Delta_i, \mu'_i - \mu^*\} \exp(-\mathbb{E}_{\nu\pi}[T_i(n)](d_i + \varepsilon)). \end{aligned}$$

Rearranging and taking the limit inferior leads to

$$\begin{aligned} \liminf_{n \rightarrow \infty} \frac{\mathbb{E}_{\nu\pi}[T_i(n)]}{\log(n)} &\geq \frac{1}{d_i + \varepsilon} \liminf_{n \rightarrow \infty} \frac{\log\left(\frac{n \min\{\Delta_i, \mu'_i - \mu^*\}}{2(R_n + R'_n)}\right)}{\log(n)} \\ &= \frac{1}{d_i + \varepsilon} \left(1 - \limsup_{n \rightarrow \infty} \frac{\log(R_n + R'_n)}{\log(n)}\right) = \frac{1}{d_i + \varepsilon}, \end{aligned}$$

where the last equality follows from the definition of consistency, which says that for any $p > 0$ there exists a constant C_p such that for sufficiently large n , $R_n + R'_n \leq C_p n^p$, which implies that

$$\limsup_{n \rightarrow \infty} \frac{\log(R_n + R'_n)}{\log(n)} \leq \limsup_{n \rightarrow \infty} \frac{p \log(n) + \log(C_p)}{\log(n)} = p,$$

which gives the result since $p > 0$ was arbitrary and by taking the limit as ε tends to zero. \square

The next theorem gives $d_{\mathcal{C}}(P, \alpha)$ for common choices of \mathcal{C} .

THEOREM 16.2 *The following hold:*

(a) *Let $\sigma^2 > 0$ and $\mathcal{C} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\}$, then*

$$d_{\mathcal{C}}(\mathcal{N}(\mu, \sigma^2), \alpha) = \frac{(\mu - \alpha)^2}{2\sigma^2}.$$

(b) *Let $\mathcal{C} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$, then*

$$d_{\mathcal{C}}(\mathcal{N}(\mu, \sigma^2), \alpha) = \frac{1}{2} \log\left(1 + \frac{(\mu - \alpha)^2}{\sigma^2}\right).$$

(c) Let $\mathcal{C} = \{\mathcal{B}(\mu) : \mu \in [0, 1]\}$, then

$$d_{\mathcal{C}}(\mathcal{B}(\mu), \alpha) = \mu \log\left(\frac{\mu}{\alpha}\right) + (1 - \mu) \log\left(\frac{1 - \mu}{1 - \alpha}\right).$$

(d) Let $\mathcal{C} = \{\mathcal{U}(a, b) : a, b \in \mathbb{R}\}$, then

$$d_{\mathcal{C}}(\mathcal{U}(a, b), \alpha) = \log\left(1 + \frac{2((a+b)/2 - \alpha)^2}{b-a}\right).$$

Proof of part (a) Fix $\sigma^2 > 0$ and note that the class $\mathcal{C} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\}$ is parameterised by the mean. Therefore for any $\mu \in \mathbb{R}$ and $\alpha > \mu$ we have

$$d_{\mathcal{C}}(\mathcal{N}(\mu, \sigma^2), \alpha) = \inf_{\theta > \alpha} D(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\theta, \sigma^2)) = \inf_{\theta > \alpha} \frac{(\mu - \theta)^2}{2\sigma^2} = \frac{(\mu - \alpha)^2}{2\sigma^2}. \quad \square$$

The reader is asked to complete the remaining parts in Exercise 16.1. It appears that the lower bound and definition of $c^*(\nu, \mathcal{E})$ are quite fundamental quantities in the sense that for most classes \mathcal{E} it appears there exists a policy π for which

$$\lim_{n \rightarrow \infty} \frac{R_n(\pi, \nu)}{\log(n)} = c^*(\nu, \mathcal{E}) \quad \text{for all } \nu \in \mathcal{E}. \quad (16.3)$$

This justifies calling a policy **asymptotically optimal** on class \mathcal{E} if Eq. (16.3) holds. For example, UCB from Chapter 8 and KL-UCB from Chapter 10 are asymptotically optimal for $\mathcal{E}_{\mathcal{N}}^K(1)$ and $\mathcal{E}_{\mathcal{B}}^K$ respectively.

16.2 Finite-time bounds

The proofs that follow use the same technique as what we already saw. For future reference we extract the common part, which summarizes what can be obtained by chaining the high-probability Pinsker inequality with the divergence decomposition lemma.

LEMMA 16.1 *Let $\nu = (P_i)$ and $\nu' = (P'_i)$ be K -action stochastic bandits that differ only in the distribution of the reward for action $i \in [K]$. Assume that i is suboptimal in ν and uniquely optimal in ν' . Let $\lambda = \mu_i(\nu') - \mu_i(\nu)$. Then for any policy π ,*

$$\mathbb{E}_{\nu\pi}[T_i(n)] \geq \frac{\log\left(\frac{\min\{\lambda - \Delta_i(\nu), \Delta_i(\nu)\}}{4}\right) + \log(n) - \log(R_n(\nu) + R_n(\nu'))}{D(P_i, P'_i)}. \quad (16.4)$$

The lemma holds for finite n and any ν and can be used to derive finite-time instance-dependent lower bounds for any environment class \mathcal{E} that is rich enough. The following result provides a finite-time instance-dependence bound for Gaussian bandits where the asymptotic notion of consistency is replaced by an assumption that the minimax regret is not too large. This assumption alone is enough to show that no policy that is remotely close to minimax optimal can be much better than UCB on any instance.

THEOREM 16.3 Let $C, p > 0$ and π be a policy such that $R_n(\pi, \nu) \leq Cn^p$ for all $\nu \in \mathcal{E}_N^K$. Then for any $\nu \in \mathcal{E}_N^K$ and $\varepsilon \in (1, 2)$ it holds that

$$R_n(\pi, \nu) \geq 2 \sum_{i: \Delta_i > 0} \left(\frac{(1-p)\log(n) + \log\left(\frac{\varepsilon\Delta_i}{8C}\right)}{\Delta_i} \right)^+, \quad (16.5)$$

where $(x)^+ = \max(x, 0)$ is the positive part of $x \in \mathbb{R}$.

Proof Let i be suboptimal in ν and choose $\nu' \in \mathcal{E}_N^K$ such that $\mu_j(\nu') = \mu_j(\nu)$ for $j \neq i$ and $\mu_j(\nu') = \mu_j(\nu) + \Delta_i(1 + \varepsilon)$. Then by Lemma 16.1 with $\lambda = \Delta_i(1 + \varepsilon)$,

$$\begin{aligned} \mathbb{E}_{\nu\pi}[T_i(n)] &\geq \frac{2}{\Delta_i^2(1+\varepsilon)^2} \left(\log\left(\frac{n}{2Cn^p}\right) + \log\left(\frac{\min\{\lambda - \Delta_i, \Delta_i\}}{4}\right) \right) \\ &= \frac{2}{\Delta_i^2(1+\varepsilon)^2} \left((1-p)\log(n) + \log\left(\frac{\varepsilon\Delta_i}{8C}\right) \right). \end{aligned}$$

Plugging this into the basic regret decomposition identity (Lemma 4.2) gives the result. \square

When $p = 1/2$ the leading term in this lower bound is approximately half that of the asymptotic bound. This effect may be real: the class of policies considered is larger than in the asymptotic lower bound and so there is the possibility that the policy that is best tuned for a given environment achieves a smaller regret.

16.3 Notes

- 1 We mentioned that for most classes \mathcal{E} there is a policy satisfying Eq. (16.3). Its form is derived from the lower bound, and by making some additional assumptions on the underlying distributions. For details, see the article by Burnetas and Katehakis [1996], which is also the original source of Theorem 16.1.
- 2 The analysis in this chapter only works for unstructured classes. Without this assumption a policy can potentially learn about the reward from one arm by playing other arms and this greatly reduces the regret. Lower bounds for structured bandits are more delicate and will be covered on a case-by-case basis in subsequent chapters.
- 3 The classes analyzed in Theorem 16.2 are all parametric, which makes the calculation possible analytically. There has been relatively little analysis in the non-parametric case, but we know of three exceptions for which we simply refer the reader to the appropriate source. The first is the class of distributions with bounded support: $\mathcal{C} = \{P : \text{Supp}(P) \subseteq [0, 1]\}$, which has been analyzed exactly [Honda and Takemura, 2010]. The second is the class of distributions with semi-bounded support, $\mathcal{C} = \{P : \text{Supp}(P) \subseteq (-\infty, 1]\}$ [Honda and Takemura, 2015]. The third is the class of distributions with bounded kurtosis, $\mathcal{C} = \{P : \text{Kurt}_{X \sim P}[X] \leq \kappa\}$. For details see Lattimore [2017].

16.4 Bibliographic remarks

Asymptotic optimality via a consistency assumption first appeared in the seminal paper by [Lai and Robbins \[1985\]](#), which was later generalized by [Burnetas and Katehakis \[1996\]](#). In terms of upper bounds, there now exist policies that are asymptotic optimal for single-parameter exponential families [[Cappé et al., 2013](#)]. Until recently, there were no results on asymptotic optimality for multi-parameter classes of reward distributions. There has been some progress on this issue recently for the Gaussian distribution with unknown mean and variance [[Cowan et al., 2015](#)] and for the uniform distribution [[Cowan and Katehakis, 2015](#)]. There are plenty of open questions related to asymptotically optimal strategies for nonparametric classes of reward distributions. When the reward distributions are discrete and finitely supported an asymptotically optimal policy is given by [Burnetas and Katehakis \[1996\]](#), though the precise constant is hard to interpret. A relatively complete solution is available for classes with bounded support [[Honda and Takemura, 2010](#)]. Already for the semi-bounded case things are getting murky [[Honda and Takemura, 2015](#)]. One of the authors thinks that classes with bounded kurtosis are quite interesting, but here things are only understood up to constant factors [[Lattimore, 2017](#)]. An asymptotic variant of Theorem 16.3 is by [Salomon et al. \[2013\]](#). Finite-time instance-dependent lower bounds have been proposed by several authors including [Kulkarni and Lugosi \[2000\]](#) for two arms and [Garivier et al. \[2016c\]](#), [Lattimore \[2018\]](#) for the general case.

16.5 Exercises

16.1 Prove parts (b), (c) and (d) of Theorem 16.2.

16.2 Let $\mathcal{R}(\mu)$ be the shifted Rademacher distribution, which for $\mu \in \mathbb{R}$ and $X \sim \mathcal{R}(\mu)$ is characterized by $\mathbb{P}(X = \mu + 1) = \mathbb{P}(X = \mu - 1) = 1/2$.

- (a) Show that $d_C(\mathcal{R}(\mu), \alpha) = \infty$ for any $\mu < \alpha$.
- (b) Design a policy π for bandits with shifted Rademacher rewards such that the regret is bounded by

$$R_n(\pi, \nu) \leq CK \quad \text{for all } n \text{ and } \nu \in \times \mathcal{C},$$

where $C > 0$ is a universal constant.

- (c) The results from parts (a) and (b) seem to contradict the heuristic analysis in Note 1 at the end of Chapter 15. Explain.

16.3 Let π be a consistent policy for a single parameter exponential family as explained in Exercise 10.4 in Chapter 10. Prove the upper bound given in part (h) is tight.

16.4 Let $\mathcal{C} = \{P : \text{there exists a } \sigma^2 \geq 0 \text{ such that } P \text{ is } \sigma^2\text{-subgaussian}\}$.

- (a) Find a distribution P such that $P \notin \mathcal{C}$.
 (b) Suppose that $P \in \mathcal{C}$ has mean $\mu \in \mathbb{R}$. Prove that $d_{\mathcal{C}}(P, \alpha) = 0$ for all $\alpha > \mu$.
 (c) Let $\mathcal{E} = \{(P_i) : P_i \in \mathcal{C} \text{ for all } 1 \leq i \leq K\}$. Prove that if $K > 1$, then for all consistent policies π ,

$$\liminf_{n \rightarrow \infty} \frac{R_n(\pi, \nu)}{\log(n)} = \infty \quad \text{for all } \nu \in \mathcal{E}.$$

- (d) Let $f : \mathbb{N} \rightarrow [0, \infty)$ be any monotone increasing function with $\lim_{n \rightarrow \infty} f(n)/\log(n) = \infty$. Prove there exists a policy π such that

$$\limsup_{n \rightarrow \infty} \frac{R_n(\pi, \nu)}{f(n)} = 0 \quad \text{for all } \nu \in \mathcal{E},$$

where \mathcal{E} is as in the previous part.

- (e) Conclude there exists a consistent policy for \mathcal{E} .

16.5 Use Lemma 16.1 to prove Theorem 15.1, possibly with different constants.

16.6 Let $K = 2$ and for $\nu \in \mathcal{E}_{\mathcal{N}}^2$ let $\Delta(\nu) = \max\{\Delta_1(\nu), \Delta_2(\nu)\}$. Suppose that π is a policy such that for all $\nu \in \mathcal{E}_{\mathcal{N}}^2$ with $\Delta(\nu) \leq 1$ it holds that

$$R_n(\pi, \nu) \leq \frac{C \log(n)}{\Delta(\nu)}. \quad (16.6)$$

- (a) Give an example of a policy satisfying Eq. (16.6).
 (b) Assume that $i = 2$ is suboptimal for ν and $\alpha \in (0, 1)$ be such that $\mathbb{E}_{\nu\pi}[T_2(n)] = \frac{1}{2\Delta(\nu)^2} \log(\alpha)$. Let ν' be the alternative environment where $\mu_1(\nu') = \mu_1(\nu)$ and $\mu_2(\nu') = \mu_1(\nu) + 2\Delta(\nu)$. Show that

$$\exp(-D(\mathbb{P}_{\nu\pi}, \mathbb{P}_{\nu'\pi})) = \frac{1}{\alpha}.$$

- (c) Let A be the event that $T_2(n) \geq n/2$. Show that

$$\mathbb{P}_{\nu\pi}(A) \leq \frac{2C \log(n)}{n\Delta^2} \quad \text{and} \quad \mathbb{P}_{\nu'\pi}(A) \geq \frac{1}{2\alpha} - \frac{2C \log(n)}{n\Delta^2}.$$

- (d) Show that

$$R_n(\pi, \nu') \geq \frac{n\Delta}{2} \left(\frac{1}{2\alpha} - \frac{2C \log(n)}{n\Delta^2} \right).$$

- (e) Show that $\alpha \geq \frac{n\Delta^2}{8C \log(n)}$ and conclude that

$$R_n(\pi, \nu) \geq \frac{1}{2\Delta(\nu)} \log \left(\frac{n\Delta^2}{8C \log(n)} \right).$$

- (f) Generalize the argument to an arbitrary number of arms.

16.7 Let $K > 1$ and $p \in [0, 1)$ and π be a policy such that for all $\mathcal{E}_{\mathcal{N}}^K$ so that for all $\nu \in \mathcal{E}_{\mathcal{N}}^K$ it holds that

$$\limsup_{n \rightarrow \infty} \frac{R_n(\pi, \nu)}{\log(n)} = \sum_{i: \Delta_i > 0} \frac{2(1+p)}{\Delta_i}.$$

Let $\hat{R}_n(\pi, \nu) = n\mu^*(\nu) - \sum_{t=1}^n \mu_{A_t}(\nu)$ be the random regret and prove that

$$\limsup_{n \rightarrow \infty} \sup_{\nu \in \mathcal{E}_n} \frac{\log(\mathbb{V}[\hat{R}_n(\pi, \nu)])}{(1-p)\log(n)} \geq 1.$$

17 High Probability Lower Bounds

The lower bounds proven in the last two chapters were for stochastic bandits. In this chapter we prove high probability lower bounds for both stochastic and adversarial bandits. Recall that for adversarial bandit $x \in [0, 1]^{nK}$ and policy π the random regret is

$$\hat{R}_n(\pi, x) = \max_{i \in [K]} \sum_{t=1}^n x_{ti} - x_{tA_t}$$

and the (expected) regret is $R_n(\pi, x) = \mathbb{E}[\hat{R}_n(\pi, x)]$. To set expectations, remember that in Chapter 12 we proved two high probability upper bounds on the regret of Exp3-IX. In the first we showed there exists a policy π such that for all adversarial bandits $x \in [0, 1]^{nK}$ and $\delta \in (0, 1)$ it holds with probability at least $1 - \delta$ that

$$\hat{R}_n(\pi, x) = O\left(\sqrt{Kn \log(K)} + \sqrt{\frac{Kn}{\log(K)}} \log\left(\frac{1}{\delta}\right)\right). \quad (17.1)$$

We also gave a version of the algorithm that depended on $\delta \in (0, 1)$ for which with probability at least $1 - \delta$,

$$\hat{R}_n(\pi, x) = O\left(\sqrt{Kn \log\left(\frac{K}{\delta}\right)}\right). \quad (17.2)$$

The important difference is the order of quantifiers. In the first we have a single algorithm and a high-probability guarantee that holds simultaneously for any confidence level. The second algorithm needs the confidence level to be specified in advance. The price for using the generic algorithm appears to be $\sqrt{\log(1/\delta)/\log(K)}$, which is usually quite small but not totally insignificant. We will see that both bounds are tight up to constant factors, which implies that knowing the desired confidence level in advance really does help. One reason why choosing the confidence level in advance is not ideal is that the resulting high-probability bound cannot be integrated to prove a bound in expectation. For algorithms satisfying (17.1) the expected regret can be bounded by

$$R_n(\pi, x) \leq \int_0^\infty \mathbb{P}(\hat{R}_n \geq x) dx = O(\sqrt{Kn \log(K)}). \quad (17.3)$$

On the other hand, if the high-probability bound only holds for a single δ as in (17.2), then it seems hard to do much better than

$$R_n \leq n\delta + O\left(\sqrt{Kn \log\left(\frac{K}{\delta}\right)}\right),$$

which with the best choice of δ leads to a bound of $O(\sqrt{Kn \log(n)})$. It turns out that this argument cannot be strengthened and algorithms with the strong high-probability regret cannot be near-optimal in expectation. For simplicity we start with the stochastic setting before explaining how to convert the arguments to the adversarial model.

17.1 Stochastic bandits

There is no randomness in the expected regret, so in order to derive a high probability bound we define the **random pseudo regret** by

$$\tilde{R}_n = \sum_{i=1}^K T_i(n) \Delta_i,$$

which is a random variable through the pull counts $T_i(n)$.



For all results in this section we let $\mathcal{E}^K \subset \mathcal{E}_{\mathcal{N}}^K$ denote the set of K -armed Gaussian bandits with suboptimality gaps bounded by one. For $\mu \in [0, 1]^d$ we let $\nu_\mu \in \mathcal{E}^K$ be the Gaussian bandit with means μ .

THEOREM 17.1 *Let $n \geq 1$ and $K \geq 2$ and $B > 0$ and π be a policy such that for any $\nu \in \mathcal{E}^K$,*

$$R_n(\pi, \nu) \leq B\sqrt{(K-1)n}. \quad (17.4)$$

Let $\delta \in (0, 1)$. Then there exists a bandit ν in \mathcal{E}^K such that

$$\mathbb{P}\left(\tilde{R}_n(\pi, \nu) \geq \frac{1}{4} \min\left\{n, \frac{1}{B} \sqrt{(K-1)n} \log\left(\frac{1}{4\delta}\right)\right\}\right) \geq \delta.$$

Proof Let $\Delta \in (0, 1/2]$ be a constant to be tuned subsequently and $\nu = \nu_\mu$ where the mean vector $\mu \in \mathbb{R}^d$ is defined by $\mu_1 = \Delta$ and $\mu_i = 0$ for $i > 1$. abbreviate $R_n = R_n(\pi, \nu)$ and $\mathbb{P} = \mathbb{P}_{\nu\pi}$ and $\mathbb{E} = \mathbb{E}_{\nu\pi}$. Let $i = \operatorname{argmin}_{i>1} \mathbb{E}[T_i(n)]$. Then by Lemma 4.2 and the assumption in Eq. (17.4),

$$\mathbb{E}[T_i(n)] \leq \frac{R_n}{\Delta} \leq \frac{B}{\Delta} \sqrt{\frac{n}{K-1}}. \quad (17.5)$$

Define alternative bandit $\nu' = \nu_{\mu'}$ where $\mu' \in \mathbb{R}^d$ is equal to μ except $\mu'_i = \mu_i + 2\Delta$. Abbreviate $\mathbb{P}' = \mathbb{P}_{\nu'\pi}$ and $\tilde{R}_n = \tilde{R}_n(\pi, \nu)$ and $\tilde{R}'_n = \tilde{R}_n(\pi, \nu')$. By Lemma 4.2

and Pinsker's inequality (Theorem 14.2) and the divergence decomposition (Lemma 15.1) we have

$$\begin{aligned} \mathbb{P}\left(\tilde{R}_n \geq \frac{\Delta n}{2}\right) + \mathbb{P}\left(\tilde{R}'_n \geq \frac{\Delta n}{2}\right) &\geq \mathbb{P}\left(T_i(n) \geq \frac{n}{2}\right) + \mathbb{P}\left(T_i(n) < \frac{n}{2}\right) \\ &\geq \frac{1}{2} \exp(-D(\mathbb{P}, \mathbb{P}')) \geq \frac{1}{2} \exp\left(-2B\Delta\sqrt{\frac{n}{K-1}}\right) \geq 2\delta, \end{aligned}$$

where the last line follows by choosing

$$\Delta = \min\left\{\frac{1}{2}, \frac{1}{2B}\sqrt{\frac{K-1}{n}} \log\left(\frac{1}{4\delta}\right)\right\}.$$

The result follows since $\max\{a, b\} \geq (a+b)/2$. \square

COROLLARY 17.1 *Let $n \geq 1$ and $K \geq 2$. Then for any policy π and $\delta \in (0, 1)$ such that*

$$n\delta \leq \sqrt{n(K-1) \log\left(\frac{1}{4\delta}\right)} \quad (17.6)$$

there exists a bandit problem $\nu \in \mathcal{E}^K$ such that

$$\mathbb{P}\left(\tilde{R}_n(\pi, \nu) \geq \frac{1}{4} \min\left\{n, \sqrt{\frac{n(K-1)}{2} \log\left(\frac{1}{4\delta}\right)}\right\}\right) \geq \delta. \quad (17.7)$$

Proof We prove the result by contradiction. Assume that the conclusion does not hold for π and let $\delta \in (0, 1)$ satisfy (17.6). Then for any bandit problem $\nu \in \mathcal{E}^K$ the expected regret of π is bounded by

$$R_n(\pi, \nu) \leq n\delta + \sqrt{\frac{n(K-1)}{2} \log\left(\frac{1}{4\delta}\right)} \leq \sqrt{2n(K-1) \log\left(\frac{1}{4\delta}\right)}.$$

Therefore π satisfies the conditions of Theorem 17.1 with $B = \sqrt{2 \log(1/(4\delta))}$, which implies that there exists some bandit problem $\nu \in \mathcal{E}^K$ such that (17.7) holds, contradicting our assumption. \square

COROLLARY 17.2 *Let $K \geq 2$ and $p \in (0, 1)$ and $B > 0$. Then there does not exist a policy π such that for all $n \geq 1$, $\delta \in (0, 1)$ and $\nu \in \mathcal{E}^K$,*

$$\mathbb{P}\left(\tilde{R}_n(\pi, \nu) \geq B\sqrt{(K-1)n} \log^p\left(\frac{1}{\delta}\right)\right) < \delta$$

Proof We proceed by contradiction. Suppose that such a policy exists. Choosing δ sufficiently small and n sufficiently large ensures that

$$\frac{1}{B} \log\left(\frac{1}{4\delta}\right) \geq B \log^p\left(\frac{1}{\delta}\right) \quad \text{and} \quad \frac{1}{B} \sqrt{n(K-1)} \log\left(\frac{1}{4\delta}\right) \leq n.$$

Now by assumption for any $\nu \in \mathcal{E}^K$ we have

$$\begin{aligned} R_n(\pi, \nu) &\leq \int_0^\infty \mathbb{P}(\tilde{R}_n(\pi, \nu) \geq x) dx \\ &\leq B\sqrt{n(K-1)} \int_0^\infty \exp(-x^{1/p}) dx \leq B\sqrt{n(K-1)}. \end{aligned}$$

Therefore by the Theorem 17.1 there exists a bandit $\nu \in \mathcal{E}^K$ such that

$$\begin{aligned} \mathbb{P}\left(\tilde{R}_n(\pi, \nu) \geq B\sqrt{n(K-1)} \log\left(\frac{1}{\delta}\right)\right) \\ \geq \mathbb{P}\left(\tilde{R}_n(\pi, \nu) \geq \frac{1}{4} \min\left\{n, \frac{1}{B}\sqrt{n(K-1)} \log\left(\frac{1}{4\delta}\right)\right\}\right) \geq \delta, \end{aligned}$$

which contradicts our assumption and completes the proof. \square



We suspect there exists a policy π and universal constant $B > 0$ such that for all $\nu \in \mathcal{E}^K$,

$$\mathbb{P}\left(\tilde{R}_n(\pi, \nu) \geq B\sqrt{Kn} \log\left(\frac{1}{\delta}\right)\right) \leq \delta.$$

Except for the issue of unbounded rewards we would have this for Exp3-IX and suspect the analysis of that algorithm could more-or-less be adapted to this setting. Care would be required to deal with the unbounded rewards, but we expect the math to go through with minor adaptations to the algorithm.

17.2 Adversarial bandits

We now explain how to translate the ideas in the previous section to the adversarial model. Throughout we assume a fixed policy π . Let $\Omega = [0, 1]^{nK}$ and let $x \in \Omega$ be an adversarial bandit environment. Recall the random regret is

$$\hat{R}_n(x) = \max_{i \in [K]} \sum_{t=1}^n (x_{ti} - x_{tA_t}),$$

where the randomness in \hat{R}_n is due to the policy only.

THEOREM 17.2 *Let $c, C > 0$ be sufficiently small/large universal constants and $K \geq 2$, $n \geq 1$ and $\delta \in (0, 1)$ be such that $n \geq CK \log(1/(2\delta))$. Then there exists a reward sequence $x \in [0, 1]^{nK}$ such that*

$$\mathbb{P}\left(\hat{R}_n(x) \geq c\sqrt{nK \log\left(\frac{1}{2\delta}\right)}\right) \geq \delta.$$

The proof is technical and messy, but also contains some nuggets of interest. For the sake of brevity we explain only the high level ideas and refer the

reader elsewhere for the gory details. There are two difficulties in translating the arguments in the previous section to the adversarial model. First, in the adversarial model we need the rewards to be bounded in $[0, 1]$. The second difficulty is we now analyse the adversarial regret rather than the random pseudo-regret.

Suppose we sample $X \in \Omega$ from distribution Q on $(\Omega, \mathcal{B}(\Omega))$ and let \mathbb{P}_Q be the distribution of $\hat{R}_n(X)$.

CLAIM 17.1 *Let \mathbb{P}_x be the distribution of $\hat{R}_n(x)$ and $u > 0$. If $\mathbb{P}_Q(\hat{R}_n(X) \geq u) \geq \delta$, then there exists an $x \in \Omega$ such that $\mathbb{P}_x(\hat{R}_n(x) \geq u)$.*

The next step is to choose Q and argue that $\mathbb{P}(\hat{R}_n(X) \geq u) \geq \delta$ for sufficiently large u . To do this we need a truncated normal distribution. Defining clipping function

$$\text{clip}_{[0,1]}(x) = \begin{cases} 1 & \text{if } x > 1 \\ 0 & \text{if } x < 0 \\ x & \text{otherwise.} \end{cases}$$

Let $\sigma, \Delta > 0$ be constants that we'll tune later and η_1, \dots, η_n a sequence of independent random variables with $\eta_t \sim \mathcal{N}(1/2, \sigma^2)$. For each $i \in [K]$ let Q_i be the distribution of $X \in \Omega$ where

$$X_{tj} = \begin{cases} \text{clip}_{[0,1]}(\eta_t + \Delta) & \text{if } j = 1 \\ \text{clip}_{[0,1]}(\eta_t + \Delta) & \text{if } j = i \text{ and } i \neq 1 \\ \text{clip}_{[0,1]}(\eta_t) & \text{otherwise,} \end{cases}$$

Notice that under any Q_i for fixed t the random variables X_{t1}, \dots, X_{tK} are not independent, but for fixed j the random variables X_{1j}, \dots, X_{tj} are independent and identically distributed. We will let the reader justify for themselves that this is equivalent to a stochastic bandit model.

CLAIM 17.2 *If $\sigma > 0$ and $\Delta = \sigma \sqrt{\frac{K-1}{2n} \log \left(\frac{1}{6\delta} \right)}$, then there exists an arm i such that*

$$\mathbb{P}_{Q_1}(T_i(n) < n/2) \geq 2\delta.$$

The proof of this claim follows along the same lines as the theorems in the previous section. All that changes is the calculation of the relative entropy. The last step is to relate $T_i(n)$ to the random regret. In the stochastic model this was straightforward, but for adversarial bandits there is an additional step. Notice that under Q_i it holds that $X_{ti} - X_{tA_t} \geq 0$ and that if $X_{ti}, X_{tA_t} \in (0, 1)$, then $X_{ti} - X_{tA_t} = \Delta$. In other words, if no clipping occurs, then $X_{ti} - X_{tA_t} = \Delta$. The following claim upper bounds the number of rounds in which clipping occurs with high probability.

CLAIM 17.3 *If $\sigma = 1/10$ and $\Delta < 1/8$ and $n \geq 32 \log(2/\delta)$, then*

$$\mathbb{P}_{Q_i} \left(\sum_{t=1}^n \mathbb{I}\{X_{ti}, X_{tA_t} \in (0, 1)\} \geq \frac{3n}{4} \right) \geq 1 - \delta.$$

By combining the first two claims with a union bound we know there exists an arm i such that

$$\mathbb{P}_{Q_i} \left(\hat{R}_n \geq \frac{n\Delta}{4} \right) \geq \delta,$$

which by the definition of Δ and Claim 17.1 implies the first part of the theorem.

17.3 Notes

- 1 The adversarial bandits used in Section 17.2 had the interesting property that the same arm has the best reward in every round (not just the best mean). It is perhaps a little surprising that algorithms cannot exploit this fact.
- 2 In Theorem 17.2 we did not make any assumptions on the algorithm. If we had assumed the algorithm enjoyed an expected regret bound of $R_n \leq B\sqrt{Kn}$, then we could conclude that for each sufficiently small $\delta \in (0, 1)$ there exists an adversarial bandit such that

$$\mathbb{P} \left(\hat{R}_n \geq \frac{c}{B} \sqrt{Kn} \log \left(\frac{1}{2\delta} \right) \right) \geq \delta,$$

which shows that our high probability upper bounds for Exp3-IX are nearly tight.

17.4 Bibliographic remarks

Though none of the results are terribly surprising, we do not know of any references except the recent paper by [Gerchinovitz and Lattimore \[2016\]](#).

17.5 Exercises

17.1 Prove each of the claims in Section 17.2.

Part V

Contextual and Linear Bandits

Suppose you want to use a bandit algorithm to decide which ads to display on your website. Can you reasonably do this with one of the models/algorithms proposed so far? A simple idea is to treat each available ad as an arm and for each arriving user choose the ad using your favourite bandit algorithm. The reward is one if they click on the ad and zero otherwise. No doubt you immediately realize that this is not exactly a perfect fit. In fact there are many problems:

- (*Delayed rewards*) You don't actually observe a user not clicking an ad. Maybe you give zero reward if the click does not happen soon, but this introduces delays and you cannot update the statistics between plays.
- (*Non-stationarity*) Unless the time period is very short, then there is significant non-stationarity in the preferences of users. None of the algorithms discussed so far behave well in this situation.
- (*Ignoring information*) If the number of available ads is large (it usually is), then treating each action independently is probably not a good idea. Users interested in ads for Mercedes might also be interested in other car ads. In many cases there is also information available about the user, such as prior purchasing decisions, and this too should be taken into account when deciding which ad is chosen to each user.

The challenge of delayed rewards and non-stationarity will be discussed elsewhere. The purpose of this part is to focus on the situation where (a) the algorithm has access to contextual information at the beginning of each round and (b) the outcome of playing one arm may yield information about other arms. For example, in the display ad problem the contextual information might consist of summary statistics of the current user and the ads in the action set might be classified into different categories. Exactly how to encode the additional structures is a non-trivial problem. As usual there is a trade-off between using models rich enough to model (almost) any world and using simple models for which generalization is easy, but for which the assumptions imposed on the world through the model are more severe.

Except for the first chapter, which is generic, the focus of this part will be on the special case that the expected reward of each arm is a linear function of some feature vector in a way that will made precise in Chapter 19. Along the way we will discuss many generalizations and give references to the literature. One aspect that will play a far larger role is computation. While finite-armed bandits with few arms present few challenges in this respect, when the number of actions is very large or the information structure of the feedback model is not so easily separable, then computation can be a serious challenge.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

18 Contextual Bandits

In many bandit problems the learner has access to additional information at the beginning of each round. Consider the problem of designing a movie recommendation system. Clearly it would be inadvisable to ignore demographic information about the user making the request, or any other contextual history such as previously watched movies or ratings. None of the algorithms presented so far can take this kind of additional information into account and the benchmark (regret) also does not measure performance relative to other sources of information. Imagine the results of trying to find the best single movie in hindsight for all users. In this chapter we will present an augmented framework and regret definition that better models the many real-world problems where contextual information is available.



Whenever you design a new benchmark, there are several factors to consider. Competing with a poor benchmark does not make sense, since even an algorithm that perfectly matches the benchmark will perform poorly. At the same time, competing with a better benchmark can be harder from a learning point of view and this penalty must be offset against the benefits.

The tradeoff just described is fundamental to all machine learning problems. In statistical estimation, the analogue tradeoff is known as the **bias-variance tradeoff**. We will not attempt to answer the question of how to resolve this tradeoff in this chapter because first we need to see how to effectively compete with improved benchmarks.

18.1 Contextual bandits: one bandit per context

In a contextual bandit problem everything works the same as in a bandit problem except the learner receives a context at the beginning of each round. The hope is that specializing the action to the context can help collect more reward. While contextual bandits can be studied in both the adversarial and stochastic frameworks, in this chapter we focus on the adversarial model. To remind the reader of the notation, let $x_{tk} \in [0, 1]$ be the reward for arm k in round t , which

is chosen in advance by the adversary. The interaction model is what you would expect:

For rounds $t = 1, 2, \dots, n$:

- 1 Learner observes context $c_t \in \mathcal{C}$ where \mathcal{C} is an arbitrary fixed set of contexts.
 - 2 Learner selects distribution P_t on $[K]$ and samples A_t from P_t .
 - 3 Learner observes reward $X_t = x_{tA_t}$.
-

A natural way to define the regret is by

$$R_n = \mathbb{E} \left[\sum_{c \in \mathcal{C}} \max_{k \in [K]} \sum_{t \in [n]: c_t = c} (x_{tk} - X_t) \right]. \quad (18.1)$$

The difference is that now we are trying to compete with the best context-dependent policy in hindsight, rather than the best fixed action. If the set of possible contexts is finite, then a simple approach is to use a separate instance of Exp3 for each context. Let

$$R_{nc} = \mathbb{E} \left[\max_{k \in [K]} \sum_{t \in [n]: c_t = c} (x_{tk} - X_t) \right]$$

be the regret due to context $c \in \mathcal{C}$. When using a separate instance of Exp3 for each context we can use the results of Chapter 11 to bound

$$R_{nc} \leq 2 \sqrt{K \sum_{t=1}^n \mathbb{I}\{c_t = c\} \log(K)}, \quad (18.2)$$

where the sum inside the square root counts the number of times context $c \in \mathcal{C}$ is observed. Because this is not known in advance it is important to use an anytime version of Exp3 for which the above regret bound holds without needing to tune a learning rate that depends on the number of times the context is observed (see Exercise 28.9). Substituting (18.2) into the regret leads to

$$R_n = \sum_{c \in \mathcal{C}} R_{nc} \leq 2 \sum_{c \in \mathcal{C}} \sqrt{K \log(K) \sum_{t=1}^n \mathbb{I}\{c_t = c\}}. \quad (18.3)$$

The magnitude of the right-hand side depends on the distribution of observed contexts. On one extreme there is only one observed context and the bound is the same as the standard finite-armed bandit problem. The other extreme occurs when all contexts are observed equally often, in which case we have

$$R_n \leq 2 \sqrt{nK|\mathcal{C}| \log(K)}. \quad (18.4)$$

Jensen's inequality applied to Eq. (18.3) shows that this really is the worst case (Exercise 18.1).



It is important to emphasize that the regret in Eq. (18.4) is different than the regret studied in Chapter 11. If we ignore the context and run the standard Exp3 algorithm, then we would have

$$\mathbb{E} \left[\sum_{t=1}^n X_t \right] \geq \max_{i \in [K]} \sum_{t=1}^n x_{ti} - 2\sqrt{Kn \log(K)}.$$

Using one version of Exp3 per context leads to

$$\mathbb{E} \left[\sum_{t=1}^n X_t \right] \geq \sum_{c \in \mathcal{C}} \max_{i \in [K]} \sum_{t \in [n]: c_t = c} x_{ti} - 2\sqrt{Kn|\mathcal{C}| \log(K)}.$$

Which of these bounds is preferable depends on the magnitude of n and how useful the context is. When n is very large the second bound is more likely to be preferable. On the other hand, the second bound is completely vacuous when $n \leq 4K|\mathcal{C}| \log(K)$.

18.2 Bandits with expert advice

For large context sets using one bandit algorithm per context will almost always be a poor choice because the additional precision is wasted unless the amount of data is enormous. Fortunately, however, it is seldom the case that the context set is both large and unstructured. To illustrate a common situation we return to the movie recommendation theme where the actions are movies and the context contains user information such as age, gender and recent movie preferences. In this case the context space is combinatorially large, but there is clearly a significant amount of structure inherited from the fact that the space of movies is highly structured and users in similar demographics are more likely to have similar preferences.

Another way to write Eq. (18.1) is to let Φ be the set of all functions from $\mathcal{C} \rightarrow [K]$. Then

$$R_n = \mathbb{E} \left[\max_{\phi \in \Phi} \sum_{t=1}^n (x_{t\phi(c_t)} - X_t) \right]. \quad (18.5)$$

The discussion above suggests we might prefer to choose Φ to be a slightly smaller set. There are many ways to do this, some of which we describe below:

Partitions

Let $\mathcal{P} \subset 2^{\mathcal{C}}$ be a partition of \mathcal{C} , which means that sets in \mathcal{P} are disjoint and $\cup_{P \in \mathcal{P}} P = \mathcal{C}$. Then define Φ to be the set of functions from \mathcal{C} to $[K]$ that are constant on each partition of \mathcal{P} . In this case we can run a version of Exp3 for each partition, which means the regret depends on the number of parts $|\mathcal{P}|$ rather than on the number of contexts.

Similarity functions

Let $s : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ be a function that we think of as measuring the similarity between pairs of contexts on the $[0, 1]$ -scale. Then let Φ be the set of functions $\phi : \mathcal{C} \rightarrow [K]$ such that the average dissimilarity

$$\frac{1}{|\mathcal{C}|^2} \sum_{c,d \in \mathcal{C}} (1 - s(c, d)) \mathbb{I}\{\phi(c) \neq \phi(d)\}$$

is below a user-tuned threshold $\theta \in (0, 1)$. It is not clear anymore that we can control the regret (18.5) using some simple meta algorithm on Exp3, but keeping the regret small is still a meaningful objective.

From supervised learning to bandits with expert advice

Yet another option is to run your favorite supervised learning method, training on batch data to find a collection of predictors $\phi_1, \dots, \phi_M : \mathcal{C} \rightarrow [K]$. Then we could use a bandit algorithm to compete with the best of these in an online fashion. This has the advantage that the offline training procedure can bring in the power of batch data and the whole army of supervised learning, without relying on potentially inaccurate evaluation methods that aim to pick the best of the pack. And why pick if one does not need to?

The possibilities are endless, but in any case, we would end up with a set of functions Φ with the goal of competing with the best of them. This suggests the idea that perhaps we should think more generally about some subset Φ of functions without necessarily considering the internal structure of Φ . This is the viewpoint that we will take. In fact, we will bring this one step further by noticing that once Φ has been chosen the contexts themselves play very little role. All we need in each round is the output of each function. This leads to a setting called **bandits with expert advice**.

In this model there are M experts. At the beginning of each round the experts announce their predictions of which actions are the most promising. For the sake of generality, we allow the experts to report not only a single prediction, but a probability distribution over the actions. The interpretation of this probability distribution is that the expert, if the decision was left to them, would choose the action for the round at random from the probability distribution it reported. As discussed before, in an adversarial setting it is natural to consider randomized algorithms, hence one should not be too surprised that the experts are also allowed to randomize. An application to an important practical problem is illustrated in Fig. 18.1.

The predictions of the M experts in round t is represented by a matrix $E^{(t)} \in [0, 1]^{M \times K}$ where the m th row $E_m^{(t)}$, a probability distribution of K , is the recommendation of expert m for round t . The learner and the environment, including the expert interact as follows:



Figure 18.1 Prediction with expert advice. The experts, upon seeing a foot give expert advice on what socks should fit it best. If the owner of the foot is happy, the recommendation system earns a cookie!

For rounds $t = 1, 2, 3, \dots, n$:

- 1 Learner observes predictions of all experts, $E^{(t)}$.
- 2 Learner selects a distribution P_t on $[K]$ in some way.
- 3 Action A_t is sampled from P_t and the reward is $X_t = x_{tA_t}$.

The regret of the learner is with respect to the total expected reward of the best *expert*:

$$R_n = \mathbb{E} \left[\max_{m \in [M]} \sum_{t=1}^n E_m^{(t)} x_t - \sum_{t=1}^n X_t \right]. \quad (18.6)$$



There is a delicate choice to be made about whether to allow the experts predictions of the experts to depend on the actions of the learner. Or whether they should be fixed from the beginning of the game in an oblivious manner. While the framework does allow learning experts, the regret definition above is not really meaningful in this case because the total reward of any of the experts will also depend on the actions chosen by the learner (through $E^{(t)}$) and in this case a more meaningful benchmark is to compare with the total reward of the experts computed under the assumption that the learner chooses some fixed action all the time. Chapter 37 will consider this type of regret for a specific problem class. However, in this chapter we restrict ourselves to non-learning, oblivious experts.

```

1: Input:  $n, K, M, \eta, \gamma$ 
2: Set  $Q_1 = (1/M, \dots, 1/M) \in [0, 1]^{1 \times M}$  (a row vector)
3: for  $t = 1, \dots, n$  do
4:   Receive advice  $E^{(t)}$ 
5:   Choose the action  $A_t \sim P_t$ , where  $P_t = Q_t E^{(t)}$ 
6:   Receive the reward  $X_t = x_{tA_t}$ 
7:   Estimate the action rewards:  $\hat{X}_{ti} = 1 - \frac{\mathbb{I}\{A_t=i\}}{P_{ti} + \gamma} (1 - X_t)$ 
8:   Propagate the rewards to the experts:  $\tilde{X}_t = E^{(t)} \hat{X}_t$ 
9:   Update the distribution  $Q_t$  using exponential weighting:

```

$$Q_{t+1,i} = \frac{\exp(\eta \tilde{X}_{ti}) Q_{ti}}{\sum_j \exp(\eta \tilde{X}_{tj}) Q_{tj}} \quad \text{for all } i \in [M]$$

```
10: end for
```

Algorithm 10: Exp4 algorithm

18.3 Can it go higher? Exp4

Exp4 is not just an increased version number, but stands for **E**xponential weighting for **x**ploration and **p**rofitation with **e**xperts. The idea of the algorithm is very simple. Since exponential weighting worked so well in the standard bandit problem we should adopt it to the problem at hand. However, since the goal is to compete with the best expert in hindsight, it is not the actions that we should score, but the experts. The algorithm maintains a probability distribution Q_t over experts and use this to come up with the next action. Once the action is chosen, we use our favorite reward estimation procedure to estimate the rewards for all the actions, which is then used to estimate how much total reward the individual experts would have made so far. The reward estimates are then used to update Q_t using exponential weighting. The pseudocode of the algorithm is given in Algorithm 10.

Note that A_t can be chosen in two steps, first sampling $M_t \in [M]$ from Q_t and then choosing $A_t \in [K]$ from $E_{M_t, \cdot}^{(t)}$. The reader can verify that (given the past) the probability distribution of the so-selected action is also P_t . The algorithm uses $O(M)$ memory and $O(MK)$ computation per round. Hence it is only practical when M and K are reasonable.

18.4 Regret analysis

We restrict our attention to the case when $\gamma = 0$, which is the original algorithm. The version where $\gamma > 0$ is called Exp4-IX and its analysis is left to the reader in Exercise 18.3.

THEOREM 18.1 *Let $\gamma = 0$ and $\eta = \sqrt{2 \log(M)/(nK)}$ and denote by R_n the*

expected regret of Exp4 defined in Algorithm 10 after n rounds. Assume that the experts are deterministic and oblivious. Then,

$$R_n \leq \sqrt{2nK \log(M)}. \quad (18.7)$$

The proof will use the following lemma:

LEMMA 18.1 *For any $m^* \in [M]$ it holds that*

$$\sum_{t=1}^n \tilde{X}_{tm^*} - \sum_{t=1}^n \sum_{m=1}^M Q_{tm} \tilde{X}_{tm} \leq \frac{\log(M)}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{m=1}^M P_{tm} (1 - \tilde{X}_{tm})^2.$$

After translating the notation, the proof of Lemma 18.1 can be extracted from the analysis of Exp3 in the proof of Theorem 11.2, a task that we leave to the reader in Exercise 18.2.

Proof of Theorem 18.1 Let m^* be the index of the best performing expert in hindsight:

$$m^* = \operatorname{argmax}_{m \in [M]} \sum_{t=1}^n E_m^{(t)} x_t. \quad (18.8)$$

Applying Lemma 18.1 shows that

$$\sum_{t=1}^n \tilde{X}_{tm^*} - \sum_{t=1}^n \sum_{m=1}^M Q_{tm} \tilde{X}_{tm} \leq \frac{\log(M)}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{m=1}^M Q_{tm} (1 - \tilde{X}_{tm})^2. \quad (18.9)$$

Let $\mathcal{F}_t = \sigma(E^{(1)}, A_1, E^{(2)}, A_2, \dots, A_{t-1}, E^{(t)})$, and introduce $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_t]$. When $\gamma = 0$ the estimator \hat{X}_{ti} is unbiased so that $\mathbb{E}_t[\hat{X}_t] = x_t$ and

$$\mathbb{E}_t[\tilde{X}_t] = \mathbb{E}_t[E^{(t)} \hat{X}_t] = E^{(t)} \mathbb{E}[\hat{X}_t] = E^{(t)} x_t. \quad (18.10)$$

Since Q_t is \mathcal{F}_t -measurable, using the tower rule for conditional expectation, taking expectations of both sides of Eq. (18.9) we get

$$R_n \leq \frac{\log(M)}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{m=1}^M \mathbb{E}[Q_{tm} (1 - \tilde{X}_{tm})^2], \quad (18.11)$$

where we also used the assumption that the experts are oblivious, and hence m^* is non-random.

Like in Chapter 11, losses are more convenient than rewards to work with. Let $\hat{Y}_{ti} = 1 - \hat{X}_{ti}$, $y_{ti} = 1 - x_{ti}$ and $\tilde{Y}_{tm} = 1 - \tilde{X}_{tm}$. Note that $\tilde{Y}_t = E^{(t)} \hat{Y}_t$ and recall also the notation $A_{ti} = \mathbb{I}\{A_t = i\}$, which means that $\hat{Y}_{ti} = \frac{A_{ti} y_{ti}}{P_{ti}}$ and

$$\mathbb{E}_t[\tilde{Y}_{tm}^2] = \mathbb{E}_t \left[\left(\frac{E_{mA_t}^{(t)} y_{tA_t}}{P_{tA_t}} \right)^2 \right] = \sum_{i=1}^K \frac{(E_{mi}^{(t)} y_{ti})^2}{P_{ti}} \leq \sum_{i=1}^K \frac{E_{mi}^{(t)}}{P_{ti}}. \quad (18.12)$$

Therefore using the definition of P_{ti} ,

$$\begin{aligned}\mathbb{E} \left[\sum_{m=1}^M Q_{tm} (1 - \tilde{X}_{tm})^2 \right] &\leq \mathbb{E} \left[\sum_{m=1}^M Q_{tm} \sum_{i=1}^K \frac{E_{mi}^{(t)}}{P_{ti}} \right] \\ &= \mathbb{E} \left[\sum_{i=1}^K \frac{\sum_{m=1}^M Q_{tm} E_{mi}^{(t)}}{P_{ti}} \right] = K.\end{aligned}$$

Substituting into Eq. (18.11) leads to

$$R_n \leq \frac{\log(M)}{\eta} + \frac{\eta n K}{2} = \sqrt{2nK \log(M)}. \quad \square$$

Let us see how this theorem can be applied to the contextual bandit where \mathcal{C} is a finite set and Φ is the set of all functions from $\mathcal{C} \rightarrow [K]$. To each of these functions $\phi \in \Phi$ we associate an expert m with $E_{mi}^{(t)} = \mathbb{I}\{\phi(c_t) = i\}$. Then $M = K^{\mathcal{C}}$ and Theorem 18.1 says that

$$R_n \leq \sqrt{2nK|\mathcal{C}| \log(K)},$$

which is the same bound we derived using an independent copy of Exp3 for each context. More generally, if \mathcal{C} is arbitrary (possibly infinite) and Φ is a finite set of functions from \mathcal{C} to $[K]$, then the theorem ensures that

$$R_n \leq \sqrt{2nK \log(|\Phi|)}.$$

These results seem quite promising already, but in fact there is another improvement possible. Define random variable E_t^* by

$$E_t^* = \sum_{s=1}^t \sum_{i=1}^K \max_{m \in [M]} E_{mi}^{(s)}.$$

By modifying the algorithm to use an adaptive learning rate of $\eta_t = \sqrt{\log(M)/E_t^*}$ one can prove the following theorem.

THEOREM 18.2 *Assume the same conditions as in Theorem 18.1, except that let $\eta_t = \sqrt{\log(M)/E_t^*}$. Then, there exists a universal constant $C > 0$ such that*

$$R_n \leq C \sqrt{E_n^* \log(M)}.$$

The proof of this result is not hard and is left to the reader in Exercise 18.4. The bound on the right-hand side of the above inequality is *data-dependent* since it depends on E_n^* . It is not hard to see (Exercise 18.7) that

$$E_n^* \leq n \min(K, M) \tag{18.13}$$

and as such this bound is much better than the bound of Theorem 18.1 when $M \leq K$. One can think of E_n^*/n as the effective number of experts which depends on the degree of disagreement, or diversity in the experts' recommendations. The bound tells us that Exp4 with the suggested learning rate is able to *adapt* to the degree of disagreement between the experts. In fact, it is reasonable that learning

becomes easier (and the regret bound will be smaller) when the experts tend to agree. At the same time, what is the use of having many experts if they tend to agree? This is another manifestation of the bias-variance tradeoff mentioned at the beginning. That Exp4 with the proposed adaptive learning rate is able to speed up learning when there is a chance to do so should be reassuring.

18.5 Notes

- 1 Perhaps the most important point of this chapter beyond the algorithms is to understand that there are tradeoffs between having a larger competitor class and a more meaningful definition of the regret that this entails. This is very similar to the tradeoff involved in considering algorithms tuned for a specific environment class (e.g., considering Bernoulli bandits, as opposed to bandits with subgaussian noise). Indeed, similarly to what happens when a smaller competitor class is chosen, a more restricted environment class usually allows faster learning, but tuning to a more restricted class runs the risk of losing on performance when the environment that the bandit algorithm runs on does not belong to the restricted class. (The lack of proved guarantees should not be mistaken for the lack of guarantees!)
- 2 The Exp4 algorithm serves as a tremendous building block for other bandit problems by defining your own experts. The best example of this is the application of Exp4 to nonstationary bandits that we explore in Chapter 31. Here, a combinatorially large set of experts is considered, and yet a fast implementation of Exp4 can be demonstrated to exist. That this is possible is more the exception than the rule. In the lack of such an efficient implementation, Exp4 can still be useful when working with a combinatorially large set of experts just to demonstrate an upper bound on the regret (for an example see Exercise 18.5).
- 3 The bandits with expert advice framework is clearly more general than contextual bandits. With the terminology of the bandits with expert advice framework, the contextual bandit problem arises when the experts are given by static $\mathcal{C} \rightarrow [K]$ maps.
- 4 A significant challenge is that a naive implementation of Exp4 has running time $O(MK)$ per round, which can be enormous if either M or K is large. In general there is no solution to this problem, but in some cases the computation can be reduced significantly. One situation where this is possible is when the learner has access to an **optimization oracle** that for any context/reward sequence that returns the expert that would collect the most reward in this sequence (this is equivalent to solving the offline problem Eq. (18.8)). In Chapter 30 we show how to use an offline optimization oracle to learn efficiently in combinatorial bandit problems. The idea is to solve a randomly perturbed optimization problem and then show that the randomness in the outputs provides sufficient exploration.

- 5 In the stochastic contextual bandit problem it is assumed that the context and reward vector form a sequence of independent and identically distributed random variables. Let Φ be a set of $\mathcal{C} \rightarrow [K]$ maps and suppose the learner has access to an optimization oracle capable of finding

$$\operatorname{argmax}_{\phi \in \Phi} \sum_{s=1}^t x_{s\phi(c_s)}$$

for any sequence of reward vectors x_1, \dots, x_t and contexts c_1, \dots, c_t . Under these circumstances there exists a polynomial-time algorithm for which the regret is essentially as the bound in Theorem 18.1.

With access to such an oracle, for stochastic contextual bandit problems there exists a polynomial-time algorithm for which the regret is essentially the same as that stated in Theorem 18.1 [Agarwal et al., 2014]. The algorithm computes importance-weighted estimates of the rewards in each round. These are used to estimate the regret of all the experts. Based on this, a distribution over the experts (with a small support) is computed by solving a feasibility problem: the distribution is constrained so that the importance weights will not be too large, while the regret estimates averaged over the chosen distribution will stay small. To reduce the computation cost, this distribution is updated periodically with the length of the interval between the updates exponentially growing. The significance of this result is that it reduces contextual bandits to (cost-sensitive) empirical risk-minimization (ERM), which means that any advance in solving cost-sensitive ERM problems automatically translates to bandits.

- 6 The development of efficient algorithms for ERM is a major topic in supervised learning. Note that ERM can be NP-hard even in simple cases like linear classification [Shalev-Shwartz and Ben-David, 2009, §8.7].
- 7 As noted earlier, the bound on the regret stated in Theorem 18.2 is data-dependent. Thinking of an instance of an adversarial bandit prediction with expert advice problem as the joint choice of the rewards (x_1, x_2, \dots) and the expert predictions $(E^{(1)}, E^{(2)}, \dots)$ we may also call the bound instance-dependent. These two expressions are in fact synonyms of each other, but the stochastic bandit literature mostly uses instance-dependent, while the adversarial online learning literature mostly uses the term data-dependent. In any case, as explained earlier, when a data, or instance-dependent bound is tight enough to imply the worst-case optimal bounds, they are preferred as they give us more information about the algorithm, or, when paired with a matching or nearly matching lower bound, about the problem class.
- 8 There are many points we have not developed in detail. One is high probability bounds, which we saw in Chapter 12 and can also be derived here. We also have not mentioned lower bounds. The degree to which the bounds are tight depends on whether or not there is additional structure in the experts. In later

chapters we will see examples where the results are essentially tight, but there are also cases where they are not.

18.6 Bibliographic remarks

For a good account on the history of contextual bandits see the article by [Tewari and Murphy \[2017\]](#). The Exp4 algorithm was introduced by [Auer et al. \[2002b\]](#) and Theorem 18.1 essentially matches Theorem 7.1 of this paper (with a slightly better constant). [McMahan and Streeter \[2009\]](#) noticed that neither the number of experts nor the size of the action set are what really matters for the regret, but rather the extent to which the experts tend to agree. [McMahan and Streeter \[2009\]](#) also introduced the idea of solving a linear program to find an exploration policy that computes a distribution over the actions such that for any action i and round t the computed probability of i is lower bound by the maximum of a constant multiple of $P_t(i)$. This is meant to ensure sufficient exploration while staying close to the output of the exponential weights distribution. The idea of explicitly optimizing a probability distribution with these objectives in mind is at the heart of several works [[Agarwal et al., 2014](#), for example]. While Theorem 18.2 is inspired by this work, the result appears to be new and goes beyond the work of [McMahan and Streeter \[2009\]](#) because it shows that all one needs is to adapt the learning rate based on the degree of agreement amongst the experts. [Neu \[2015a\]](#) proves high probability bounds for Exp4-IX. You can follow in his footsteps by solving Exercise 18.3. Another way to get high probability bounds is to generalize Exp3.P, which was done by [Beygelzimer et al. \[2011\]](#). As we mentioned in Item 5, there exist efficient algorithms for stochastic contextual bandit problems when a suitable optimization oracle is available [[Agarwal et al., 2014](#)]. An earlier attempt to address the problem of reducing contextual bandits to cost-sensitive ERM is by [Dudik et al. \[2011\]](#). The adversarial case of static experts is considered by [Syrkanis et al. \[2016\]](#) who prove suboptimal (worse than \sqrt{n}) regret bounds under various conditions for follow the perturbed leader for the transductive setting when the contexts are available at the start. The case when the contexts are independent and identically distributed, but the reward is adversarial has been studied by [Lazaric and Munos \[2009\]](#) for the finite expert case, while [Rakhlin and Sridharan \[2016\]](#) considered the case when an ERM oracle is available. The paper of [Rakhlin and Sridharan \[2016\]](#) also considers the more realistic case when only an approximation oracle is available for the ERM problem. What is notable about this work is they demonstrate regret bounds with a moderate blow-up, but without changing the definition. [Kakade et al. \[2008\]](#) consider contextual bandit problems with adversarial context-loss sequences, where all but one action suffers a loss of one in every round. This can also be seen as an instance of **multiclass classification with bandit feedback** where labels to be predicted are identified with actions and the only feedback received is whether the label predicted was correct, with the goal of making as few mistakes as possible. Since

minimizing the regret is in general hard in this non-convex setting, just like most of the machine learning literature on classification, [Kakade et al. \[2008\]](#) provide results in the form of mistake bounds for linear classifiers where the baseline is not the number of mistakes of the best linear classifier, but is a convex upper bound on it. The recent book by [Shalev-Shwartz and Ben-David \[2009\]](#) lists some hardness results for ERM. For a more comprehensive treatment, the reader can consult the book by [Kearns and Vazirani \[1994\]](#).

18.7 Exercises

18.1 Let \mathcal{C} be a finite context set and let $c_1, \dots, c_n \in \mathcal{C}$ be an arbitrary sequence of contexts.

- (a) Show that $\sum_{c \in \mathcal{C}} \sqrt{\sum_{t=1}^n \mathbb{I}\{c_t = c\}} \leq \sqrt{n|\mathcal{C}|}$.
- (b) Assume that n is an integer multiple of $|\mathcal{C}|$. Show that the choice that maximizes the right-hand side of the previous inequality is the one when each context occurs $n/|\mathcal{C}|$ times.

18.2 Prove Lemma 18.1.

18.3 In this exercise you will prove an analogue of Theorem 12.1 for Exp4-IX. In the contextual setting the random regret is

$$\hat{R}_n = \max_{m \in [M]} \sum_{t=1}^n \left(E_m^{(t)} x_t - X_t \right).$$

Design an algorithm accepting parameter $\delta \in (0, 1)$ such that

$$\mathbb{P} \left(\hat{R}_n \geq C \left(\sqrt{nK \log(K)} + \sqrt{\frac{nK}{\log(K)}} \log \left(\frac{1}{\delta} \right) \right) \right) \leq \delta.$$

18.4 Prove Theorem 18.2.

18.5 Let x_1, \dots, x_n be a sequence of reward vectors chosen in advance by an adversary with $x_t \in [0, 1]^K$. Furthermore, let o_1, \dots, o_n be a sequence of observations, also chosen in advance by an adversary with $o_t \in [O]$ for some fixed $O \in \mathbb{N}^+$. Then let \mathcal{H} be the set of functions $\phi : [O]^m \rightarrow [K]$ where $m \in \mathbb{N}^+$. In each round the learner observes o_t should choose an action A_t based on $o_1, A_1, X_1, \dots, o_{t-1}, A_{t-1}, X_{t-1}, o_t$ and the regret is

$$R_n = \min_{\phi \in \mathcal{H}} \sum_{t=1}^n x_{tA_t} - x_{t\phi(o_t, o_{t-1}, \dots, o_{t-m})},$$

where $o_t = 1$ for $t \leq 0$. This means the learner is competing with the best

predictor in hindsight that uses only the last m observations. Prove there exists an algorithm such that

$$\mathbb{E}[R_n] \leq \sqrt{2nmK \log(O)}.$$

18.6 In this problem we consider non-oblivious experts. Consider the following modified regret definition:

$$R'_n = \max_{m \in [M]} \mathbb{E} \left[\sum_{t=1}^n E_m^{(t)} x_t - \sum_{t=1}^n X_t \right].$$

Show that:

- (a) $R'_n \leq R_n$ regardless of whether the experts are oblivious or not.
- (b) Theorem 18.1 remains valid for non-oblivious experts if in Eq. (18.7) we replace R_n with R'_n . In particular, explain how to modify the proof.
- (c) Research question: Give a non-trivial bound on R_n .

18.7 Prove Eq. (18.13).

18.8 [The epoch-greedy algorithm, [Langford and Zhang, 2008](#)] Consider a stochastic contextual bandit environment, where the context-reward pairs (C_t, X_t) form an i.i.d. sequence, with $C_t \in \mathcal{C}$ and $X_t \in [0, 1]^K$. Let $\Phi \subset \{\phi : \phi : \mathcal{C} \rightarrow [K]\}$ be a set of static experts and assume that we have access to an oracle $\mathcal{O}(x, c)$ that can compute $\text{argmax}_{\phi \in \Phi} \sum_{s=1}^t x_{t,\phi(c_t)}$ for any $x = (x_s)_s$, $c = (c_s)_s$ sequences of reward-vectors and contexts $(x_s \in \mathbb{R}^K, c_s \in \mathcal{C})$.

The **epoch-greedy algorithm** works in phases of length $1 < \tau_1 < \tau_2 < \dots$ of increasing length. In the first round of phase $m = 1, 2, \dots$, the algorithm receives context \tilde{C}_m and then performs an exploration step: An action $\tilde{A}_m \in [K]$ is chosen uniformly at random. Let \tilde{X}_m denote the reward received in response. Next, the algorithm constructs the reward estimates $\hat{X}_{m,k} = \frac{1}{K} \mathbb{I}\{\tilde{A}_m = k\} \tilde{X}_m$ and finds the expert ϕ_m whose usage so far would have incurred the most total reward: $\phi_m = \text{argmax}_{\phi \in \Phi} \sum_{p=1}^m \hat{X}_{p,\phi(\tilde{C}_p)}$. In the remaining $\tau_m - 1$ rounds of phase m , the advice of ϕ_m is followed: Upon receiving context C_t in round t (during this phase), action $A_t = \phi_m(C_t)$ is used.

- (a) Let Φ be finite. Show that with an appropriate choice of $(\tau_m)_m$, the expected regret R_n of epoch-greedy after n steps is $R_n \leq O(n^{2/3} |\Phi|^{1/3})$.
- (b) Extension to VC-dimension!
- (c) Can the result be extended to the case when the context sequence $(c_t)_t$ is an arbitrary fixed sequence, but $X_t \sim P_{c_t}$ for some family $(P_c)_c$ of distributions?

18.9 [Experimenting] Different exploration strategies? McMahan-Streeter...

18.10 [Application of Exp4] Fill this in..

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

19 Stochastic Linear Bandits

Contextual bandits generalize the finite-armed setting by allowing the learner to make use of side information. This chapter focusses on a specific type of contextual bandit problem in the stochastic setup where the reward is assumed to have a linear structure that allows for learning to transfer from one context to another. This leads to a useful and rich model that will be the topic of the next few chapters. To begin we describe the **stochastic linear bandit** problem and start the process of generalizing the upper confidence bound algorithm.

19.1 Stochastic contextual bandits

The stochastic contextual bandit problem mirrors the adversarial contextual bandit setup discussed in Chapter 18. At the beginning of round t the learner observes a context $C_t \in \mathcal{C}$, which may be random or not. Having observed the context, the learner chooses their action $A_t \in [K]$ based on the information available. So far everything is the same as the adversarial setting. The difference comes from the assumption that the reward X_t satisfies

$$X_t = r(C_t, A_t) + \eta_t,$$

where $r : \mathcal{C} \times [K] \rightarrow \mathbb{R}$ is called the **reward function** and η_t is the noise, which we will assume is conditionally 1-subgaussian. Precisely, let

$$\mathcal{F}_t = \sigma(C_1, A_1, X_1, \dots, C_{t-1}, A_{t-1}, X_{t-1}, C_t, A_t)$$

be the σ -field summarizing the information available just before X_t is observed. Then we assume that

$$\mathbb{E} [\exp(\lambda \eta_t) | \mathcal{F}_t] \leq \exp\left(\frac{\lambda^2}{2}\right) \quad \text{almost surely}.$$

The noise could have been chosen to be σ -subgaussian for any known σ^2 , but like in earlier chapters we save ourselves some ink by fixing its value to $\sigma^2 = 1$. Remember from Chapter 5 that subgaussian random variables have zero mean, so the assumption also implies that $\mathbb{E}[\eta_t | \mathcal{F}_t] = 0$ and $\mathbb{E}[X_t | \mathcal{F}_t] = r(C_t, A_t)$.

If r was given, then the action in round t with the largest expected return is $A_t^* \in \operatorname{argmax}_{a \in [K]} r(C_t, a)$. Notice that this action is now a random variable

because it depends on the context C_t . The loss due to the lack of knowledge of r makes the learner incur the (expected) regret

$$R_n = \mathbb{E} \left[\sum_{t=1}^n \max_{a \in [K]} r(C_t, a) - \sum_{t=1}^n X_t \right].$$

Like in the adversarial setting, there is one big caveat in this definition of the regret. Since we did not make any restrictions on how the contexts are chosen, it could be that choosing a low-rewarding action in the first round might change the contexts observed in subsequent rounds. Then the learner could potentially achieve an even higher cumulative reward by choosing a ‘suboptimal’ arm initially. As a consequence, this definition of the regret is most meaningful when the actions of the learner do not greatly affect subsequent contexts.

One way to eventually learn an optimal policy is to estimate $r(c, a)$ for each $(c, a) \in \mathcal{C} \times [K]$ pair. As in the adversarial setting, this is ineffective when the number of context-action pairs is large. In particular, the worst-case regret over all possible contextual problems with M contexts and mean reward in $[0, 1]$ is at least $\Omega(\sqrt{nMK})$. While this may not look bad, M is often exponentially large (for example, 2^{100}). The argument for proving such worst-case lower bounds relies on designing a problem where knowledge of $r(c, \cdot)$ for context c provides no useful information about $r(c', \cdot)$ for some different context c' . Fortunately, in most interesting applications the set of contexts is highly structured, which can often be captured by some kind of smoothness of $r(\cdot, \cdot)$.

A very simple idea is to assume the learner has access to a map $\psi : \mathcal{C} \times [K] \rightarrow \mathbb{R}^d$ and that there exists an unknown parameter vector $\theta_* \in \mathbb{R}^d$ such that

$$r(c, a) = \langle \psi(c, a), \theta_* \rangle, \quad \forall (c, a) \in \mathcal{C} \times [K]. \quad (19.1)$$

The map ψ is called a **feature-map**, which is the standard nomenclature in machine learning. The idea of feature maps is best illustrated with an example. Suppose the context denotes the visitor of a website selling books, the actions are books to recommend and the reward is the revenue on a book sold. The features could indicate the interests of the visitors as well as the domain and topic of the book. If the visitors and books are assigned to finitely many categories, indicator variables of all possible combinations of these categories could be used to create the feature map. Of course, many other possibilities exist. For example you can train a neural network (deep or not) on historical data to predict the revenue and use the nonlinear map that we obtained by removing the last layer of the neural network. The subspace Ψ spanned by the **feature vectors** $\{\psi(c, a)\}_{c,a}$ in \mathbb{R}^d is called the **feature-space**.

If $\|\cdot\|$ is a norm on \mathbb{R}^d , then an assumption on $\|\theta_*\|$ encodes **smoothness** of r . In particular, from Hölder’s inequality,

$$|r(c, a) - r(c', a')| \leq \|\theta_*\| \|\psi(c, a) - \psi(c', a')\|_*,$$

where $\|\cdot\|_*$ denotes the dual of $\|\cdot\|$. Restrictions on $\|\theta_*\|$ have a similar effect to assuming that the dimensionality d . In fact, one may push this to the extreme

and allow d to be infinite, an approach which can buy tremendous flexibility and makes the linearity assumption less limiting.

19.2 Stochastic linear bandits

Stochastic linear bandits arise from realizing that when the reward is given by Eq. (19.1), then the identity of the actions becomes secondary. All that matters is the feature vector that results from choosing a given action. This justifies studying the following simplified model: In round t , the learner is given the decision set $\mathcal{A}_t \subset \mathbb{R}^d$ from which it chooses its action $A_t \in \mathcal{A}_t$ and receives reward

$$X_t = \langle A_t, \theta_* \rangle + \eta_t,$$

where η_t is 1-subgaussian given $\mathcal{A}_1, A_1, X_1, \dots, \mathcal{A}_{t-1}, A_{t-1}, X_{t-1}, \mathcal{A}_t$ and A_t . The random regret and regret are defined by

$$\begin{aligned}\hat{R}_n &= \sum_{t=1}^n \max_{a \in \mathcal{A}_t} \langle a, \theta_* \rangle - \sum_{t=1}^n X_t. \\ R_n &= \mathbb{E} [\hat{R}_n] = \mathbb{E} \left[\sum_{t=1}^n \max_{a \in \mathcal{A}_t} \langle a, \theta_* \rangle - \sum_{t=1}^n X_t \right].\end{aligned}$$

Different choices of \mathcal{A}_t lead to different settings, some of which we have seen before. For example, if $(e_i)_i$ are the unit vectors and $\mathcal{A}_t = \{e_1, \dots, e_d\}$, then the resulting stochastic linear bandit problem reduces to the finite-armed setting. On the other hand, if $\mathcal{A}_t = \{\psi(C_t, k) : k \in [K]\}$, then we have a contextual linear bandit. Yet another possibility is a **combinatorial action set** like $\mathcal{A}_t \subseteq \{0, 1\}^d$. Many combinatorial problems (such as matching, least-cost problems in directed graphs and choosing spanning trees) can be written as linear optimization problems over some combinatorial set \mathcal{A} obtained from considering incidence vectors often associated with some graph. Some of these topics will be covered later in Chapter 30.

As we have seen in earlier chapters, the UCB algorithm is an attractive approach for finite-action stochastic bandits. Its best variants are nearly minimax optimal, instance optimal and exactly optimal asymptotically. With these merits in mind, it seems quite natural to try and generalize the idea to the linear setting.

The generalization is based on the view that UCB implements the ‘optimism in the face of uncertainty’ principle, which is to act in each round as if the environment is as nice as plausibly possible. In finite-action stochastic bandits this means choosing the action with the largest upper confidence bound. In the case of linear bandits the idea remains the same, but the form of the confidence bound is more complicated because rewards received yield information about more than just the arm played.

The first step is to construct a confidence set $\mathcal{C}_t \subset \mathbb{R}^d$ based on $(A_1, X_1, \dots, A_{t-1}, X_{t-1})$ that contains the unknown parameter vector θ_* with

high probability. Leaving the details of how the confidence set is constructed aside for a moment and assuming that the confidence set indeed contains θ_* , then for any given action $a \in \mathbb{R}^d$,

$$\text{UCB}_t(a) = \max_{\theta \in \mathcal{C}_t} \langle a, \theta \rangle \quad (19.2)$$

will be an upper bound on the mean payoff of a , which is $\langle a, \theta_* \rangle$. The UCB algorithm that uses the confidence set \mathcal{C}_t at time t then selects

$$A_t = \operatorname{argmax}_{a \in \mathcal{A}_t} \text{UCB}_t(a). \quad (19.3)$$

Depending on the authors, UCB applied to linear bandits is known by many names, including LinRel (**L**inear **R**einforcement **L**earning), LinUCB and OFUL (**O**ptimism in the **F**ace of **U**ncertainty for **L**inear bandits).

The main question is how to choose the confidence set $\mathcal{C}_t \subset \mathbb{R}^d$. As usual, there are conflicting desirable properties:

- (a) \mathcal{C}_t should contain θ_* with high probability.
- (b) \mathcal{C}_t should be as small as possible.

At first sight it is not at all obvious what \mathcal{C}_t should look like. After all, it is a subset of \mathbb{R}^d , not just an interval like the confidence intervals about the empirical estimate of the mean reward for a single action that we saw in the previous chapters. We will leave the details to the next chapter, but sketch the basic approach here. Following the idea for UCB, we need an analogue for the empirical estimate of the unknown quantity, which in this case is θ^* . There are several principles one might use for deriving such an estimate. For now we use the **regularized least-squares estimator**, which is

$$\hat{\theta}_t = \operatorname{argmin}_{\theta \in \mathbb{R}^d} \left(\sum_{s=1}^t (X_s - \langle A_s, \theta \rangle)^2 + \lambda \|\theta\|_2^2 \right), \quad (19.4)$$

where $\lambda \geq 0$ is called the **penalty factor**. Choosing $\lambda > 0$ helps because it ensures that the loss function has a unique minimizer even when A_1, \dots, A_t do not span \mathbb{R}^d , which simplifies the math. The solution to Eq. (19.4) is obtained easily by differentiation and is

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^t A_s X_s, \quad (19.5)$$

where V_t is a $d \times d$ matrices given by

$$V_0 = \lambda I \quad \text{and} \quad V_t = V_0 + \sum_{s=1}^t A_s A_s^\top.$$

The matrix $V_t - V_0$ is called the **Grammian** while V_t is sometimes called the **regularized Grammian**. So $\hat{\theta}_t$ is an estimate of θ_* , which makes it natural to

choose \mathcal{C}_t to be centered at $\hat{\theta}_{t-1}$. For what follows we will simply assume that the confidence set \mathcal{C}_t is closed and satisfies

$$\mathcal{C}_t \subseteq \mathcal{E}_t = \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}}^2 \leq \beta_t \right\}, \quad (19.6)$$

where $(\beta_t)_t$ is a sequence of monotone nondecreasing constants with $\beta_1 \geq 1$ and for positive-definite matrix $A \in \mathbb{R}^{d \times d}$ and $x \in \mathbb{R}^d$ we define $\|x\|_A^2 = x^\top A x$, a notation which is justified by the fact that $\|\cdot\|_A$ is indeed a norm. The set \mathcal{E}_t is an ellipsoid centered at $\hat{\theta}_{t-1}$ and with principle axis being the eigenvectors of V_t with corresponding lengths being the reciprocal of the eigenvalues. Notice that as t grows the matrix V_t has increasing eigenvalues, which means the volume of the ellipse is also shrinking (at least, provided β_t does not grow too fast). In the next chapter we will see that $\mathcal{C}_t = \mathcal{E}_t$ is a natural choice for carefully chosen β_t , but for the rest of this chapter we will simply examine the consequence of using a confidence set satisfying Eq. (19.6) and assume all the desirable properties.

19.3 Regret analysis

We prove a regret bound for LinUCB under the assumption that the confidence intervals indeed contain the true parameter with high probability and boundedness conditions on the action set and rewards.

ASSUMPTION 19.1 The following hold:

- (a) $|\langle a, \theta_* \rangle| \leq 1$ for any $a \in \cup_t \mathcal{A}_t$.
- (b) For any $a \in \cup_t \mathcal{A}_t$, $\|a\|_2 \leq L$.
- (c) There exists a $\delta \in (0, 1)$ such that with probability $1 - \delta$, for all $t \in [n]$, $\theta_* \in \mathcal{C}_t$ where \mathcal{C}_t satisfies Eq. (19.6).

THEOREM 19.1 *Under the conditions of Assumption 19.1 with probability $1 - \delta$ the regret of LinUCB satisfies*

$$\hat{R}_n \leq \sqrt{8n\beta_n \log \left(\frac{\det V_n}{\det V_0} \right)} \leq \sqrt{8dn\beta_n \log \left(\frac{\text{trace}(V_0) + nL^2}{d \det^{\frac{1}{d}}(V_0)} \right)}.$$

Provided that β_n has polylogarithmic growth, then $\hat{R}_n = \tilde{O}(\sqrt{n})$, which matches the worst-case rate for finite-armed bandits except for logarithmic factors. We can also get a bound on the (expected) regret R_n if $\delta \leq c/\sqrt{n}$ and by combining the theorem with trivial fact that $\hat{R}_n \leq 2n$, which follows from our assumption that the magnitude of the immediate reward is bounded by one. The proof of Theorem 19.1 depends on the following lemma.

LEMMA 19.1 *Let V_0 be positive definite and $v_0 = \text{trace}(V_0)$ and $x_1, \dots, x_n \in \mathbb{R}^d$*

be a sequence of vectors with $\|x_t\|_2 \leq L < \infty$ for all $t \in [n]$. Then

$$\sum_{t=1}^n \left(1 \wedge \|x_t\|_{V_{t-1}}^2 \right) \leq 2 \log \left(\frac{\det V_n}{\det V_0} \right) \leq 2d \log \left(\frac{v_0 + nL^2}{d \det^{1/d}(V_0)} \right).$$

Proof Using that for any $u \in [0, 1]$, $u \wedge 1 \leq 2 \ln(1+u)$, we get

$$\sum_{t=1}^n \left(1 \wedge \|x_t\|_{V_{t-1}}^2 \right) \leq 2 \sum_t \log \left(1 + \|x_t\|_{V_{t-1}}^2 \right).$$

We now argue that this last expression is $\log \left(\frac{\det V_n}{\det V_0} \right)$. For $t \geq 1$ we have

$$V_t = V_{t-1} + x_t x_t^\top = V_{t-1}^{1/2} (I + V_{t-1}^{-1/2} x_t x_t^\top V_{t-1}^{-1/2}) V_{t-1}^{1/2}.$$

Hence

$$\det(V_t) = \det(V_{t-1}) \det \left(I + V_{t-1}^{-1/2} x_t x_t^\top V_{t-1}^{-1/2} \right) = \det(V_{t-1}) \left(1 + \|x_t\|_{V_{t-1}}^2 \right),$$

where the second equality follows because the matrix $I + yy^\top$ has eigenvalues $1 + \|y\|_2^2$ and 1 as well as the fact that the determinant of a matrix is the product of its eigenvalues. Putting things together we see that

$$\det(V_n) = \det(V_0) \prod_{t=1}^n \left(1 + \|x_t\|_{V_{t-1}}^2 \right),$$

which is equivalent to the first inequality that we wanted to prove. To get the second inequality note that by the inequality of arithmetic and geometric means,

$$\det(V_n) = \prod_{i=1}^d \lambda_i \leq \left(\frac{1}{d} \text{trace } V_n \right)^d \leq \left(\frac{v_0 + nL^2}{d} \right)^d,$$

where $\lambda_1, \dots, \lambda_d$ are the eigenvalues of V_n . \square

Proof of Theorem 19.1 By part (c) of Assumption 19.1 it suffices to prove the bound on the event that $\theta_* \in \mathcal{C}_t$ for all rounds $t \in [n]$. Let $A_t^* = \operatorname{argmax}_{a \in \mathcal{A}_t} \langle a, \theta_* \rangle$ be an optimal action for round t and r_t be the instantaneous regret in round t defined by

$$r_t = \langle A_t^* - A_t, \theta_* \rangle.$$

Let $\tilde{\theta}_t \in \mathcal{C}_t$ be the parameter in the confidence set for which $\langle A_t, \tilde{\theta}_t \rangle = \text{UCB}_t(A_t)$. Then using the fact that $\theta_* \in \mathcal{C}_t$ and the definition of the algorithm leads to

$$\langle A_t^*, \theta_* \rangle \leq \text{UCB}_t(A_t^*) \leq \text{UCB}_t(A_t) = \langle A_t, \tilde{\theta}_t \rangle.$$

Using Cauchy-Schwartz inequality and the assumption that $\theta_* \in \mathcal{C}_t$ and facts that $\tilde{\theta}_t \in \mathcal{C}_t$ and $\mathcal{C}_t \subseteq \mathcal{E}_t$ leads to

$$r_t = \langle A_t^* - A_t, \theta_* \rangle \leq \langle A_t, \tilde{\theta}_t - \theta_* \rangle \leq \|A_t\|_{V_{t-1}} \|\tilde{\theta}_t - \theta_*\|_{V_{t-1}} \leq 2 \|A_t\|_{V_{t-1}} \beta_t.$$

By part (a) we also have $r_t \leq 2$, which combined with $\beta_n \geq \max\{1, \beta_t\}$ yields

$$r_t \leq 2 \wedge 2\sqrt{\beta_t} \|A_t\|_{V_{t-1}^{-1}} \leq 2\sqrt{\beta_n}(1 \wedge \|A_t\|_{V_{t-1}^{-1}}).$$

Jensen's inequality shows that

$$\hat{R}_n = \sum_{t=1}^n r_t \leq \sqrt{n \sum_{t=1}^n r_t^2} \leq 2\sqrt{n\beta_n \sum_{t=1}^n (1 \wedge \|A_t\|_{V_{t-1}^{-1}}^2)}.$$

The result is completed using Lemma 19.1, which depends on part (b) of Assumption 19.1. \square

19.4 Notes

- 1 It was mentioned that ψ may map its arguments to an infinite dimensional space. There are several issues that arise in this setting. The first is whether or not the algorithm can be computed efficiently, which is usually tackled via the **kernel trick**, which assumes the existence of an efficiently computable **kernel function** $\kappa : (\mathcal{C} \times [K]) \times (\mathcal{C} \times [K]) \rightarrow \mathbb{R}$ such that

$$\langle \psi(c, a), \psi(c', a') \rangle = \kappa((c, a), (c', a')).$$

Then all operations are written in terms of the kernel function so that $\psi(c, a)$ never needs to be computed or stored. The second issue is that the statement of Theorem 19.1 depends on the dimension d and becomes vacuous when d is large or infinite. This dependence arises from Lemma 19.1, which must be replaced with a data-dependent quantity that measures the ‘effective dimension’ of the image of the data under ϕ . The final challenge is to define an appropriate confidence set. These issues have not yet been resolved in a complete way. See the bibliographic remarks for further references.

- 2 The bound given in Theorem 19.1 is essentially a worst-case style of bound, with little dependence on the parameter θ_* or the geometry of the action-set. Instance-dependent bounds for linear bandits are still an open topic of research, and the asymptotics are only understood in the special case where the action set is finite and unchanging (Chapter 25).
- 3 An obvious question is whether or not the optimization problem in Eq. (19.3) can be solved efficiently. First note that the computation of A_t can also be written as

$$(A_t, \tilde{\theta}_t) = \operatorname{argmax}_{(a, \theta) \in \mathcal{A}_t \times \mathcal{C}_t} \langle a, \theta \rangle. \quad (19.7)$$

This is a bilinear optimization problem over the set $\mathcal{A}_t \times \mathcal{C}_t$. In general, nothing much can be said about the computational efficiency of solving this problem. There are two notable special cases:

- (c) If the linear optimization problem $\max_{a \in \mathcal{A}_t} \langle a, \theta \rangle$ can be efficiently solved for any θ and \mathcal{C}_t is the convex hull of a small number of vertices: $\mathcal{C}_t = \text{co}(c_{t1}, \dots, c_{tp})$. Then it is easy to verify that the solution to Eq. (19.7) has the form (a, c_{ti}) for some $i \in [p]$. Hence the solution may be found by solving $\max_{a \in \mathcal{A}_t} \langle a, c_{t1} \rangle, \dots, \max_{a \in \mathcal{A}_t} \langle a, c_{tp} \rangle$.
- (c) If $\mathcal{C}_t = \mathcal{E}_t$ is the ellipsoid given in Eq. (19.6) and \mathcal{A}_t is a small finite set. Then the action A_t from Eq. (19.7) can be found using

$$A_t = \operatorname{argmax}_a \langle a, \hat{\theta}_t \rangle + \sqrt{\beta_t} \|a\|_{V_{t-1}^{-1}}, \quad (19.8)$$

which may be solved by simply iterating over the arms and calculating the term inside the argmax.

- 4 The previous note highlights the fact that the algorithm presented in this section has more than just a passing resemblance to the UCB algorithm introduced in earlier chapters on finite-armed bandits. The term $\langle a, \hat{\theta}_t \rangle$ may be interpreted as an empirical estimate of the reward from choosing action a and $\sqrt{\beta_t} \|a\|_{V_{t-1}^{-1}}$ is a bonus term that ensures sufficient exploration. If the penalty term vanishes ($\lambda = 0$) and $\mathcal{A}_t = \{e_1, \dots, e_d\}$ for all $t \in [n]$, then $\hat{\theta}_i$ becomes the empirical mean of action e_i and the matrix V_t is diagonal with its i diagonal entry being the number of times action e_i is used up to and including round t . Then the bonus term has order

$$\sqrt{\beta_t} \|e_i\|_{V_{t-1}^{-1}} = \sqrt{\frac{\beta_t}{T_i(t-1)}},$$

where $T_i(t-1)$ is the number of times action e_i has been chosen before the t th round. So UCB for finite-armed bandits is recovered by choosing $\beta_t = 2 \log(\cdot)$, where the term inside the logarithm can be chosen in a variety of ways as discussed in earlier chapters. Notice now that the simple analysis given in this chapter leads to a regret bound of $O(\sqrt{dn \log(\cdot)})$, which is quite close to the highly specialized analysis given in Chapters 7 to 9.

- 5 A practical extension of the linear model is the **generalized linear model** where the reward is

$$X_t = g^{-1}(\langle A_t, \theta_* \rangle + \eta_t), \quad (19.9)$$

where $g : \mathbb{R} \rightarrow \mathbb{R}$ is called the **link function**. A common choice is $g(p) = \log(p/(1-p))$, which yields the sigmoid function as the inverse: $g^{-1}(x) = 1/(1 + \exp(-x))$. Bandits with rewards from a generalized linear model have been studied by Filippi et al. [2010], who prove a bound with a similar form as Theorem 19.1. Unfortunately, however, the bound depends in a slightly unpleasant manner on the form of the link function and it seems there may be significant room for improvement.

19.5 Bibliographic remarks

Stochastic linear bandits were introduced by [Abe and Long \[1999\]](#). The first paper to consider algorithms based on the optimism principle for linear bandits is by [Auer \[2002\]](#), who considered the case when the number of actions is finite. The core ideas of the analysis of optimistic algorithms (and more) is already present in this paper. An algorithm based on confidence ellipsoids is described in the papers by [Dani et al. \[2008\]](#), [Rusmevichientong and Tsitsiklis \[2010\]](#), [Abbasi-yadkori et al. \[2011\]](#). The regret analysis presented here and the discussion of the computational questions is largely based on the former of these works, which also stresses that an expected regret of $\tilde{O}(d\sqrt{n})$ can be achieved regardless of the shape of the decision sets \mathcal{A}_t as long as the means are guaranteed to lie in a bounded interval. [Rusmevichientong and Tsitsiklis \[2010\]](#) consider both optimistic and explore-then-commit strategies which they call “phased exploration and greedy exploitation” (PEGE). They focus on the case where \mathcal{A}_t is the unit ball and show that PEGE is optimal up to logarithmic factors. The observation that explore-then-commit works for the unit ball (and other action sets with a smooth boundary) was independently made by [Abbasi-Yadkori et al. \[2009\]](#), [Abbasi-Yadkori \[2009a\]](#). Generalized linear models are credited to [Nelder and Wedderburn \[1972\]](#). We mentioned already that LinUCB was generalized to this model by [Filippi et al. \[2010\]](#). A more computationally efficient algorithm has recently been proposed by [Jun et al. \[2017\]](#). Nonlinear structured bandits where the payoff function belongs to a known set has also been studied [[Anantharam et al., 1987](#), [Russo and Roy, 2013](#), [Lattimore and Munos, 2014](#)]. The kernelized version of UCB is by [Valko et al. \[2013b\]](#). We mentioned early in the chapter that making assumptions on the norm θ_* is related to smoothness of the reward function with smoother functions leading to stronger guarantees. For an example of where this is done see the paper on ‘spectral bandits’ by [Valko et al. \[2014\]](#).

19.6 Exercises

19.1 Prove that the solution given in Eq. (19.5) is indeed the minimizer of Eq. (19.4).

19.2 Let $V_0 = \lambda I$ and $x_1, \dots, x_n \in \mathbb{R}^d$ be a sequence of vectors with $\|x_t\|_2 \leq L$ for all $t \in [n]$. Then let $V_t = V_0 + \sum_{s=1}^t x_s x_s^\top$ and show that the number of times $\|x_t\|_{V_{t-1}^{-1}} \geq 1$ is at most

$$\frac{3d}{\log(2)} \log \left(1 + \frac{L^2}{\lambda \log(2)} \right).$$



The proof of Theorem 19.1 depended on part (a) of Assumption 19.1, which asserts that the mean rewards are bounded by 1. Suppose we replace this assumption with the relaxation that there exists a $B > 0$ such that

$$\max_{t \in [n]} \sup_{a, b \in \mathcal{A}_t} \langle a - b, \theta_* \rangle \leq B.$$

Then the previous exercise allows you to bound the number of rounds when $\|x_t\|_{V_{t-1}^{-1}} \geq 1$ and in these rounds the naive bound of $r_t \leq B$ is used. For the remaining rounds the analysis of Theorem 19.1 goes through unaltered. As a consequence we see that the dependence on B is an additive constant term that does not grow with the horizon.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

20 Confidence Bounds for Least Squares Estimators

In the last chapter we derived a regret bound for a version of the upper confidence bound algorithm that depended on a particular kind of confidence set. The purpose of this chapter is to justify these choices.

Suppose that at the end of round t a bandit algorithm has chosen actions $A_1, \dots, A_t \in \mathbb{R}^d$ and received the respective payoffs X_1, \dots, X_t . Recall from the previous chapter that the **penalized least-squares** (or **ridge regression**) estimate of θ_* is the minimizer of the penalized squared empirical loss,

$$L_t(\theta) = \sum_{s=1}^t (X_s - \langle A_s, \theta \rangle)^2 + \lambda \|\theta\|_2^2,$$

where $\lambda \geq 0$ is the penalty factor. This is minimized by

$$\hat{\theta}_t = V_t(\lambda)^{-1} \sum_{s=1}^t X_s A_s \quad \text{with } V_t(\lambda) = \lambda I + \sum_{s=1}^t A_s A_s^\top. \quad (20.1)$$

It is convenient for the remainder to abbreviate $V_t = V_t(0)$.

Designing a confidence set about $\hat{\theta}_t$ when A_1, \dots, A_t have been chosen by a bandit algorithm is a surprisingly delicate matter. The difficulty stems from the fact that the actions $(A_s)_{s < t}$ are neither fixed nor independent, but are intricately correlated via the rewards. We spend the first section of this chapter building intuition by making some simplifying assumptions. Eager readers may skip directly to Section 20.1. For the rest of this section we assume that:

- 1 *Nonsingular Grammian*: $\lambda = 0$ and V_t is invertible.
- 2 *Independent subgaussian noise*: $(\eta_s)_s$ are independent and 1-subgaussian.
- 3 *Fixed design*: A_1, \dots, A_t are deterministically chosen without the knowledge of X_1, \dots, X_t .

None of these assumptions is plausible in the bandit setting, but the simplification eases the analysis and provides insight. To emphasize that A_1, \dots, A_t are deterministic we use a_s in place of A_s so that

$$V_t = \sum_{s=1}^t a_s a_s^\top \quad \text{and} \quad \hat{\theta}_t = V_t^{-1} \sum_{s=1}^t a_s X_s.$$

Note that for V_t to be non-singular it is necessary that the actions $(a_s)_{s=1}^t$ span

\mathbb{R}^d , which of course implies that $t \geq d$. Comparing θ_* and $\hat{\theta}_t$ in the direction $x \in \mathbb{R}^d$ we have

$$\begin{aligned}\langle x, \hat{\theta}_t - \theta_* \rangle &= \left\langle x, V_t^{-1} \sum_{s=1}^t a_s X_s - \theta_* \right\rangle = \left\langle x, V_t^{-1} \sum_{s=1}^t a_s (a_s^\top \theta_* + \eta_s) - \theta_* \right\rangle \\ &= \left\langle x, V_t^{-1} \sum_{s=1}^t \eta_s a_s \right\rangle = \sum_{s=1}^t \langle x, V_t^{-1} a_s \rangle \eta_s.\end{aligned}$$

Since $(\eta_s)_s$ are independent and 1-subgaussian, by Lemma 5.2 and Theorem 5.1,

$$\mathbb{P} \left(\langle x, \hat{\theta}_t - \theta_* \rangle \geq \sqrt{2 \sum_{s=1}^t \langle x, V_t^{-1} a_s \rangle^2 \log \left(\frac{1}{\delta} \right)} \right) \leq \delta.$$

A little linear algebra shows that $\sum_{s=1}^t \langle x, V_t^{-1} a_s \rangle^2 = \|x\|_{V_t^{-1}}^2$, which means that

$$\mathbb{P} \left(\langle x, \hat{\theta}_t - \theta_* \rangle \geq \sqrt{2 \|x\|_{V_t^{-1}}^2 \log \left(\frac{1}{\delta} \right)} \right) \leq \delta. \quad (20.2)$$

The next step is to convert the above bound on $\langle x, \hat{\theta}_t - \theta_* \rangle$ to a bound on $\|\hat{\theta}_t - \theta^*\|_{V_t}$. To begin this process notice that

$$\|\hat{\theta}_t - \theta^*\|_{V_t} = \langle V_t^{1/2} X, \hat{\theta}_t - \theta_* \rangle, \text{ where } X = \frac{V_t^{1/2} (\hat{\theta}_t - \theta_*)}{\|\hat{\theta}_t - \theta_*\|_{V_t}}.$$

The problem is that X is random while we have only proven (20.2) for deterministic x . The standard way of addressing problems like this is to use a **covering argument**. First we identify a finite set $\mathcal{C}_\varepsilon \subset \mathbb{R}^d$ such that whatever value X takes, there exists some $x \in \mathcal{C}_\varepsilon$ that is ε -close to X . Then a union bound and a triangle inequality allows one to finish. By its definition we have $\|X\|_2^2 = X^\top X = 1$, which means that $X \in S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. Using that $X \in S^{d-1}$ we see it suffices to “cover” S^{d-1} . For this we have the following result:

LEMMA 20.1 *There exists a set $\mathcal{C}_\varepsilon \subset \mathbb{R}^d$ with $|\mathcal{C}_\varepsilon| \leq (3/\varepsilon)^d$ such that for all $x \in S^{d-1}$ there exists a $y \in \mathcal{C}_\varepsilon$ with $\|x - y\|_2 \leq \varepsilon$.*

The proof of this lemma requires a bit work, but nothing really deep is needed. This work is deferred to the exercises Exercises 20.1 and 20.2.

Letting \mathcal{C}_ε be the covering set given by the lemma and applying a union bound and Eq. (20.2) shows that

$$\mathbb{P} \left(\text{exists } x \in \mathcal{C}_\varepsilon : \left\langle V_t^{1/2} x, \hat{\theta}_t - \theta_* \right\rangle \geq \sqrt{2 \|x\|_{V_t^{-1}}^2 \log \left(\frac{|\mathcal{C}_\varepsilon|}{\delta} \right)} \right) \leq \delta.$$

Then assuming the event inside the probability does not occur and using Cauchy-Schwarz inequality,

$$\begin{aligned} \|\hat{\theta}_t - \theta_*\|_{V_t} &= \max_{x \in S^{d-1}} \left\langle V_t^{1/2} x, \hat{\theta}_t - \theta_* \right\rangle \\ &= \max_{x \in S^{d-1}} \min_{y \in \mathcal{C}_\varepsilon} \left[\left\langle V_t^{1/2}(x - y), \hat{\theta}_t - \theta_* \right\rangle + \left\langle V_t^{1/2} y, \hat{\theta}_t - \theta_* \right\rangle \right] \\ &< \max_{x \in S^{d-1}} \min_{y \in \mathcal{C}_\varepsilon} \left[\|\hat{\theta}_t - \theta_*\|_{V_t} \|x - y\|_2 + \sqrt{2\|V_t^{1/2} y\|_{V_t^{-1}}^2 \log\left(\frac{|\mathcal{C}_\varepsilon|}{\delta}\right)} \right] \\ &\leq \varepsilon \|\hat{\theta}_t - \theta_*\|_{V_t} + \sqrt{2 \log\left(\frac{|\mathcal{C}_\varepsilon|}{\delta}\right)}. \end{aligned}$$

Rearranging yields

$$\|\hat{\theta}_t - \theta_*\|_{V_t} \leq \frac{1}{1 - \varepsilon} \sqrt{2 \log\left(\frac{|\mathcal{C}_\varepsilon|}{\delta}\right)}.$$

And now there is a tension in the choice of $\varepsilon > 0$. The term in the denominator suggests that ε should be small, but by Lemma 20.1 the cardinality of \mathcal{C}_ε grows rapidly as ε tends to zero. By lazily choosing $\varepsilon = 1/2$,

$$\mathbb{P}\left(\|\hat{\theta}_t - \theta_*\|_{V_t} \geq 2\sqrt{2\left(d \log(6) + \log\left(\frac{1}{\delta}\right)\right)}\right) \leq \delta. \quad (20.3)$$

Except for constants and other minor differences, this turns out to be about as good as you can get. Unfortunately, however, this analysis only works because V_t was assumed to be deterministic. In the active case, where A_1, \dots, A_n are chosen by a bandit algorithm, this assumption does not hold and the ideas need to be modified.

20.1 Martingale noise and Laplace's method

We now remove all of the limiting assumptions in the previous section. Of course we still need some conditions on the noise. In particular, we assume that η_1, \dots, η_t are conditionally 1-subgaussian:

$$\mathbb{E}[\exp(\alpha \eta_s) | \eta_1, \dots, \eta_{s-1}] \leq \exp\left(\frac{\alpha^2}{2}\right), \quad \text{for all } \alpha \in \mathbb{R} \text{ and } s \in [t]. \quad (20.4)$$

We have now dropped the assumption that A_1, A_2, \dots are fixed in advance and so return to the usual capitalization. We also allow arbitrary penalty factors $\lambda > 0$ and relax the assumption that V_t be invertible (though $V_t(\lambda)$ is now invertible because $\lambda > 0$). Can we still get a confidence set like what appears in (20.3)? Before diving in, we need to introduce another concept from probability theory.

DEFINITION 20.1 (Martingale difference process) Let $\mathbb{F} = (\mathcal{F}_s)_s$ be a filtration

over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A sequence of random variables $(U_s)_s$ is an \mathbb{F} -adapted martingale difference process if for all s it holds that $\mathbb{E}[U_s]$ exists and U_s is \mathcal{F}_s -measurable and $\mathbb{E}[U_s | \mathcal{F}_{s-1}] = 0$.

As usual, the filtration is often not explicitly mentioned if it is obvious from the context. The name is justified by the fact that if $(U_s)_s$ is a martingale difference process, then the partial sums $M_t = \sum_{s=1}^t U_s$ define a martingale. A more descriptive and informal name for a martingale difference process is **martingale noise**.

In the linear bandit model $(\eta_s)_s$ is martingale noise with the filtration given by $\mathcal{F}_s = \{A_1, X_1, \dots, A_{s-1}, X_{s-1}, A_s\}$. Note the inclusion of A_s in the definition of \mathcal{F}_s . The martingale noise assumption allows the noise η_s to depend on past choices, including the most recent action. This is often essential. For example, if the rewards are Bernoulli. Let us return to the construction of confidence sets. Since we want exponentially decaying tail probabilities one is tempted to try Chernoff's method:

$$\mathbb{P}\left(\|\hat{\theta}_t - \theta_*\|_{V_t} \geq u\right) \leq \inf_{\lambda > 0} \mathbb{E}\left[\exp\left(\lambda\|\hat{\theta}_t - \theta_*\|_{V_t} - \lambda u\right)\right].$$

Sadly, we do not know how to bound this expectation. Can we still somehow use Chernoff's method? Let $S_t = \sum_{s=1}^t \eta_s A_s$ and apply the ‘linearization trick’ to show that

$$\frac{1}{2}\|\hat{\theta}_t - \theta_*\|_{V_t}^2 = \max_{x \in \mathbb{R}^d} \left(\langle x, S_t \rangle - \frac{1}{2}\|x\|_{V_t}^2 \right).$$

The exponential of the term inside the maximum is a supermartingale.

LEMMA 20.2 *For all $x \in \mathbb{R}^d$ the process $M_t(x) = \exp(\langle x, S_t \rangle - \frac{1}{2}\|x\|_{V_t}^2)$ is a \mathbb{F} -adapted supermartingale.*

Proof of Lemma 20.2 That $M_t(x)$ is \mathcal{F}_t -measurable for all t is immediate from the definition. We need to show that $\mathbb{E}[M_t(x) | \mathcal{F}_{t-1}] \leq M_{t-1}(x)$ almost surely. The fact that (η_s) is martingale noise with respect to filtration (\mathcal{F}_s) means that

$$\mathbb{E}[\exp(\eta_s \langle x, A_s \rangle) | \mathcal{F}_s] \leq \exp\left(\frac{\langle x, A_s \rangle^2}{2}\right) = \exp\left(\frac{\|x\|_{A_s A_s^\top}^2}{2}\right) \quad \text{a.s.}$$

Hence

$$\begin{aligned} \mathbb{E}[M_t(x) | \mathcal{F}_{t-1}] &= \mathbb{E}\left[\exp\left(\langle x, S_t \rangle - \frac{1}{2}\|x\|_{V_t}^2\right) | \mathcal{F}_{t-1}\right] \\ &= M_{t-1}(x) \mathbb{E}\left[\exp\left(\eta_t \langle x, A_t \rangle - \frac{1}{2}\|x\|_{A_t A_t^\top}^2\right) | \mathcal{F}_{t-1}\right] \\ &\leq M_{t-1}(x) \quad \text{a.s.} \end{aligned} \quad \square$$

Combining the lemma and the linearization idea almost works. Chernoff's

method leads to

$$\begin{aligned}
 \mathbb{P}\left(\frac{1}{2}\|\hat{\theta}_t - \theta_*\|_{V_t}^2 \geq \log(1/\delta)\right) &= \mathbb{P}\left(\exp\left(\max_{x \in \mathbb{R}^d} \langle x, S_t \rangle - \frac{1}{2}\|x\|_{V_t}^2\right) \geq 1/\delta\right) \\
 &\leq \delta \mathbb{E}\left[\exp\left(\max_{x \in \mathbb{R}^d} \langle x, S_t \rangle - \frac{1}{2}\|x\|_{V_t}^2\right)\right] \\
 &= \delta \mathbb{E}\left[\max_{x \in \mathbb{R}^d} M_t(x)\right]. \tag{20.5}
 \end{aligned}$$

Now Lemma 20.2 shows that $\mathbb{E}[M_t(x)] \leq 1$. This seems quite promising, but the presence of the maximum is a setback because Jensen's inequality implies that $\mathbb{E}[\max_{x \in \mathbb{R}^d} M_t(x)] \geq \max_{x \in \mathbb{R}^d} \mathbb{E}[M_t(x)]$, which is the wrong direction to be used above. This means we cannot directly use the lemma to bound Eq. (20.5). There are two natural ways to attack this problem. The first idea is to define a finite covering set $\mathcal{C}_\varepsilon \subset \mathbb{R}^d$ so that

$$\begin{aligned}
 \mathbb{E}\left[\max_{x \in \mathbb{R}^d} M_t(x)\right] &= \mathbb{E}\left[\max_{x \in \mathbb{R}^d} \min_{y \in \mathcal{C}_\varepsilon} M_t(x) - M_t(y) + M_t(y)\right] \\
 &\leq \mathbb{E}\left[\max_{x \in \mathbb{R}^d} \min_{y \in \mathcal{C}_\varepsilon} |M_t(x) - M_t(y)|\right] + \mathbb{E}\left[\max_{y \in \mathcal{C}_\varepsilon} M_t(y)\right] \\
 &\leq \mathbb{E}\left[\max_{x \in \mathbb{R}^d} \min_{y \in \mathcal{C}_\varepsilon} |M_t(x) - M_t(y)|\right] + \sum_{y \in \mathcal{C}_\varepsilon} \mathbb{E}[M_t(y)] \\
 &\leq \varepsilon + |\mathcal{C}_\varepsilon|. \tag{20.6}
 \end{aligned}$$

The last inequality follows from a careful choice of \mathcal{C}_ε , and as usual the size of the covering set must be balanced against the required accuracy. Choosing \mathcal{C}_ε is quite non-trivial because $M_t(x) - M_t(y)$ is random, even for fixed x and y . We leave the ‘last few steps’ as an exercise (see Exercise 20.3). The second approach actually does not require us to bound Eq. (20.5), but uses it for inspiration when combined with Laplace’s method for approximating integrals of well-behaved exponentials.

Laplace's method (†)

We briefly review Laplace’s method for one-dimension functions. Assume that $f : [a, b] \rightarrow \mathbb{R}$ is twice differentiable and has a unique maximum at $x_0 \in (a, b)$ with $-q = f''(x_0) < 0$. Laplace’s method for approximating $f(x_0)$ is to compute the integral

$$I_s = \int_a^b \exp(sf(x))dx$$

for some large value of $s > 0$. From a Taylor expansion we may write

$$f(x) = f(x_0) - \frac{q}{2}(x - x_0)^2 + R(x),$$

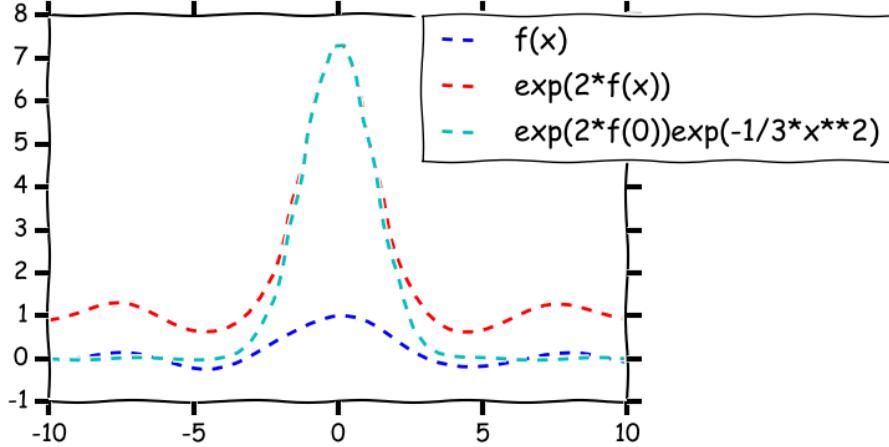


Figure 20.1 Laplace's method

where $R(x) = o((x - x_0)^2)$. Under appropriate technical assumptions,

$$I_s \sim \exp(sf(x_0)) \int_a^b \exp\left(-\frac{sq(x-x_0)^2}{2}\right) dx \quad \text{as } s \rightarrow \infty.$$

Furthermore, as s gets large

$$\int_a^b \exp\left(-\frac{sq(x-x_0)^2}{2}\right) dx \sim \int_{-\infty}^{\infty} \exp\left(-\frac{sq(x-x_0)^2}{2}\right) dx = \sqrt{\frac{2\pi}{sq}}$$

and hence

$$I_s \sim \exp(sf(x_0)) \sqrt{\frac{2\pi}{sq}}.$$

Intuitively, the dominant term in the integral I_s is $\exp(sf(x_0))$. It should also be clear that the fact that we integrate with respect to the Lebesgue measure does not matter much. We could have integrated with respect to any other measure as long as that measure puts a positive mass on the neighborhood of the maximizer. The method is illustrated on the figure shown below. The take home message of this is that if we integrate the exponential of a function that has a pronounced maximum then we can expect that the integral will be close to the exponential function of the maximum.

Method of mixtures

Laplace's approximation suggests that

$$\max_x M_t(x) \approx \int_{\mathbb{R}^d} M_t(x) dh(x), \quad (20.7)$$

where h is some measure on \mathbb{R}^d chosen so that the integral can be calculated in closed form. This is not a requirement of the method, but it does make the

argument shorter. The main benefit of replacing the maximum with an integral is that we obtain the following lemma, which you will prove in Exercise 20.4.

LEMMA 20.3 *Let h be a probability measure on \mathbb{R}^d , then $\bar{M}_t = \int_{\mathbb{R}^d} M_t(x) dh(x)$ is a \mathbb{F} -adapted supermartingale.*

THEOREM 20.1 *For all $\lambda > 0$ and $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\text{exists } t \leq n : \|S_t\|_{V_t(\lambda)^{-1}}^2 \geq 2 \log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t(\lambda))}{\lambda^d}\right)\right) \leq \delta.$$

Proof Let $H = \lambda I$ and $h = \mathcal{N}(0, H)$ and

$$\begin{aligned}\bar{M}_t &= \int_{\mathbb{R}^d} M_t(x) dh(x) \\ &= \frac{1}{\sqrt{(2\pi)^d \det(H^{-1})}} \int_{\mathbb{R}^d} \exp\left(\langle x, S_t \rangle - \frac{1}{2}\|x\|_{V_t}^2 - \frac{1}{2}\|x\|_H^2\right) dx.\end{aligned}$$

Completing the square,

$$\langle x, S_t \rangle - \frac{1}{2}\|x\|_{V_t}^2 - \frac{1}{2}\|x\|_H^2 = \frac{1}{2}\|S_t\|_{(H+V_t)^{-1}}^2 - \frac{1}{2}\|x - (H+V_t)^{-1}S_t\|_{H+V_t}^2.$$

The first term $\|S_t\|_{(H+V_t)^{-1}}^2$ does not depend on x and can be moved outside the integral, which leaves a quadratic ‘Gaussian’ term that may be integrated exactly and results in

$$\bar{M}_t = \left(\frac{\det(H)}{\det(H+V)}\right)^{1/2} \exp\left(\frac{1}{2}\|S_t\|_{(H+V_t)^{-1}}^2\right). \quad (20.8)$$

And things have worked out beautifully. Since \bar{M}_t is a nonnegative supermartingale, the maximal inequality (Theorem 3.5) shows that

$$\mathbb{P}\left(\sup_{t \in \{0, \dots, n\}} \log(\bar{M}_t) \geq \log\left(\frac{1}{\delta}\right)\right) = \mathbb{P}\left(\sup_{t \in \{0, \dots, n\}} \bar{M}_t \geq \frac{1}{\delta}\right) \leq \delta.$$

The result follows by substituting Eq. (20.8) into the above display and rearranging. \square

THEOREM 20.2 *Assuming $\delta \in (0, 1)$, then with probability at least $1 - \delta$ it holds that for all $t \in \{0, 1, \dots, n\}$,*

$$\|\hat{\theta}_t - \theta_*\|_{V_t(\lambda)} < \sqrt{\lambda}\|\theta_*\| + \sqrt{2 \log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det V_t(\lambda)}{\lambda^d}\right)}.$$

Furthermore, if $\|\theta_*\| \leq S$, then $\mathbb{P}(\text{exists } t \in [n] : \theta_* \notin \mathcal{C}_t) \leq \delta$ with

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta}_{t-1} - \theta\|_{V_{t-1}(\lambda)} \leq \sqrt{\lambda}S + \sqrt{2 \log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det V_{t-1}(\lambda)}{\lambda^d}\right)} \right\}.$$

Proof We only have to compare $\|S_t\|_{V_t(\lambda)^{-1}}$ and $\|\hat{\theta}_t - \theta_*\|_{V_t(\lambda)}$.

$$\begin{aligned}\|\hat{\theta}_t - \theta_*\|_{V_t(\lambda)} &= \|V_t(\lambda)^{-1}S_t + (V_t(\lambda)^{-1}V_t - I)\theta_*\|_{V_t(\lambda)} \\ &= \|S_t\|_{V_t(\lambda)^{-1}} + (\theta_*^\top(V_t(\lambda)^{-1}V_t - I)V_t(\lambda)(V_t(\lambda)^{-1}V_t - I)\theta_*)^{1/2} \\ &= \|S_t\|_{V_t(\lambda)^{-1}} + \lambda^{1/2}(\theta_*^\top(I - V_t(\lambda)^{-1}V_t)\theta_*)^{1/2} \\ &\leq \|S_t\|_{V_t(\lambda)^{-1}} + \lambda^{1/2}\|\theta_*\|.\end{aligned}$$

And the result follows from Theorem 20.1. \square

20.2 Notes

- 1 An alternative to the 2-norm based construction is to use 1-norms. In the fixed design setting, under the independent Gaussian noise assumption, using Chernoff's method this leads to

$$\mathcal{C}_{t+1} = \left\{ \theta \in \mathbb{R}^d : \|V^{1/2}(\hat{\theta}_t - \theta)\|_1 \leq \sqrt{2 \log(2)d^2 + 2d \log(1/\delta)} \right\}. \quad (20.9)$$

- 2 Supermartingales come up all the time in proofs relying on Chernoff's method. Just one example is the proof of Lemma 12.1. One could rewrite most the proofs involving sums of random variables relying on Chernoff's method in a way that it would become clear that proof hinges on the supermartingale property of an appropriate sequence.

20.3 Bibliographic remarks

Laplace's method is also called the 'Method of Mixtures' [Peña et al., 2008] and its use goes back to the work of Robbins and Siegmund [1970]. In practice, the improvement that results from using Laplace's method as compared to the previous ellipsoidal constructions that are based on covering arguments is quite large. A historical account of martingale methods in sequential analysis is by Lai [2009]. A simple proof of Lemma 20.1 appears as Lemma 2.5 in the book by van de Geer [2000]. Calculating covering numbers (or related packing numbers) is a whole field by itself, with open questions even in the most obvious examples. The main reference is by Rogers [1964], which by now is a little old, but still interesting.

20.4 Exercises

For Exercise 20.2 where we ask you to prove Lemma 20.1 a few standard definitions will be useful.

DEFINITION 20.2 (Covering and Packing) Let $\mathcal{A} \subset \mathbb{R}^d$. A subset $\mathcal{C} \subset \mathcal{A}$ is said to be an **ε -cover** of \mathcal{A} if $\mathcal{A} \subset \cup_{x \in \mathcal{C}} B(x, \varepsilon)$, where $B(x, \varepsilon) = \{y \in \mathbb{R}^d : \|x - y\| \leq \varepsilon\}$ is the ε ball centered at x . An **ε -packing** of \mathcal{A} is a subset $\mathcal{P} \subset \mathcal{A}$ such that for any $x, y \in \mathcal{P}$, $\|x - y\| > \varepsilon$ (note the strict inequality). The **ε -covering number** of \mathcal{A} is $N(\mathcal{A}, \varepsilon) = \min\{|\mathcal{C}| : \mathcal{C} \text{ is an } \varepsilon\text{-covering of } \mathcal{A}\}$, while the **ε -packing number** of \mathcal{A} is $M(\mathcal{A}, \varepsilon) = \max\{|\mathcal{P}| : \mathcal{P} \text{ is an } \varepsilon\text{-packing of } \mathcal{A}\}$, where we allow for both the covering and packing numbers to take on the value of $+\infty$.

 There are various generalizations of these definitions, which do not change their essence. For one, the definitions can be repeated for arbitrarily pseudo-metric spaces (instead of \mathbb{R}^d with the Euclidean distance, we can consider a set X with a $d : X \times X \rightarrow [0, \infty)$ function on it which is symmetric and satisfies the triangle inequality and that $d(x, x) = 0$ for any $x \in X$). The basic results concerning covering and packing stated in the next exercise remain valid with this more general definition. In applications we often need the logarithm of the covering and packing numbers, which are then given the new name the set's **metric entropy** (at a scale ε). As we shall see these are often close no matter whether we consider packing or covering.

We separate a useful set of a results concerning packing and covering:

20.1 [Coverings and Packings] Let $\mathcal{A} \subset \mathbb{R}^d$, B the unit ball of \mathbb{R}^d , $\text{vol}(\cdot)$ the usual volume (measure under the Lebesgue measure). For brevity let $N(\varepsilon) = N(\mathcal{A}, \varepsilon)$ and $M(\varepsilon) = M(\mathcal{A}, \varepsilon)$. Show that the following hold:

- (a) $\varepsilon \rightarrow N(\varepsilon)$ is increasing as $\varepsilon \geq 0$ is decreasing.
- (b) $M(2\varepsilon) \leq N(\varepsilon) \leq M(\varepsilon)$.
- (c) We have

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\text{vol}(\mathcal{A})}{\text{vol}(B)} \leq N(\varepsilon) \leq M(\varepsilon) \leq \frac{\text{vol}(\mathcal{A} + \frac{\varepsilon}{2}B)}{\text{vol}(\frac{\varepsilon}{2}B)} \stackrel{(*)}{\leq} \frac{\text{vol}(\frac{3}{2}\mathcal{A})}{\text{vol}(\frac{\varepsilon}{2}B)} \leq \left(\frac{3}{\varepsilon}\right)^d \frac{\text{vol}(\mathcal{A})}{\text{vol}(B)},$$

where $(*)$ holds under the assumption that $\varepsilon B \subset \mathcal{A}$ and that \mathcal{A} is convex and for $U, V \subset \mathbb{R}^d$, $c \in \mathbb{R}$, $U + V = \{u + v : u \in U, v \in V\}$ and $cU = \{cu : u \in U\}$;

- (d) Fix $\varepsilon > 0$. Then $N(\varepsilon) < +\infty$ if and only if \mathcal{A} is bounded. The same holds for $M(\varepsilon)$.

20.2 Use the results of the previous exercise to prove Lemma 20.1.

20.3 Complete the steps to show Eq. (20.6).

20.4 Prove Lemma 20.3.

20.5 [Hoeffding–Azuma] Let X_1, \dots, X_n be a sequence of random variables adapted to filtration $\mathbb{F} = (\mathcal{F}_t)_t$. Suppose that $|X_t| \in [a_t, b_t]$ almost surely for

arbitrary fixed sequences (a_t) and (b_t) with $a_t \leq b_t$ for all $t \in [n]$. Show that for any $\varepsilon > 0$,

$$\mathbb{P} \left(\sum_{t=1}^n (X_t - \mathbb{E}[X_t | \mathcal{F}_t]) \geq \varepsilon \right) \leq \exp \left(-\frac{2n^2 \varepsilon^2}{\sum_{t=1}^n (b_t - a_t)^2} \right).$$



It may help to recall Hoeffding's lemma from Note 7 in Chapter 5, which states that for random variable $X \in [a, b]$ the moment generating function satisfies

$$M_X(\lambda) \leq \exp(\lambda^2(b-a)^2/8).$$

20.6 The following simple extension of Hoeffding–Azuma is often useful. Let $n \in \mathbb{N}^+$ and (a_t) and (b_t) be fixed sequences with $a_t \leq b_t$ for all $t \in [n]$. Let X_1, \dots, X_n be a sequence of random variables adapted to filtration $\mathbb{F} = (\mathcal{F}_t)_t$ and A be an event. Assume that $\mathbb{P}(\text{exists } t \in [n] : A \text{ and } X_t \notin [a_t, b_t]) = 0$ and $\varepsilon > 0$ and show that

- (a) $\mathbb{P} \left(A \cap \sum_{t=1}^n (X_t - \mathbb{E}[X_t | \mathcal{F}_t]) \geq \varepsilon \right) \leq \exp \left(-\frac{2n^2 \varepsilon^2}{\sum_{t=1}^n (b_t - a_t)^2} \right).$
- (b) $\mathbb{P} \left(\sum_{t=1}^n (X_t - \mathbb{E}[X_t | \mathcal{F}_t]) \geq \varepsilon \right) \leq \mathbb{P}(A^c) + \exp \left(-\frac{2n^2 \varepsilon^2}{\sum_{t=1}^n (b_t - a_t)^2} \right).$



The utility of this result comes from the fact that very often the range of some adapted sequence is itself random and could be arbitrarily large with low probability (when A does not hold). A reference for the above result is the survey by McDiarmid [1998].

20.7 Let $\mathbb{F} = (\mathcal{F}_t)_{t=0}^n$ be a filtration and X_1, X_2, \dots, X_n be a sequence of \mathbb{F} -adapted random variables with $X_t \in \{-1, 0, 1\}$ and $\mu_t = \mathbb{E}[X_t | \mathcal{F}_{t-1}, X_t \neq 0]$, which we define to be zero whenever $\mathbb{P}(X_t \neq 0 | \mathcal{F}_{t-1}) = 0$. Then with $S_t = \sum_{s=1}^t (X_s - \mu_s | X_s|)$ and $N_t = \sum_{s=1}^t |X_s|$,

$$\mathbb{P} \left(\text{exists } t \leq n : |S_t| \geq \sqrt{2N_t \log \left(\frac{c\sqrt{N_t}}{\delta} \right)} \text{ and } N_t > 0 \right) \leq \delta,$$

where $c > 0$ is a universal constant.



This result appeared in a paper by the authors and others with the constant $c = 4\sqrt{2/\pi}/\text{erf}(\sqrt{2}) \approx 3.43$ [Lattimore et al., 2018].

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

21 Optimal Design for Least Squares Estimators

In the preceding chapters we introduced the linear bandit and showed how to construct confidence intervals for least squares estimators. We now study the problem of choosing actions for which these confidence intervals are small, which plays an important role in the analysis of stochastic linear bandits with finitely many arms (Chapter 22) or adversarial linear bandits (Part VI).

Let η_1, \dots, η_n be a sequence of independent 1-subgaussian random variables and $A_1, \dots, A_n \in \mathbb{R}^d$ be a fixed sequence with $\text{span}(A_1, \dots, A_n) = \mathbb{R}^d$ and Y_1, \dots, Y_n be given by $Y_t = \langle A_t, \theta_* \rangle + \eta_t$ for some $\theta_* \in \mathbb{R}^d$. Recall from the previous chapter that the least square estimator is $\hat{\theta} = V^{-1} \sum_{t=1}^n A_t Y_t$ with $V = \sum_{t=1}^n A_t A_t^\top$ the design matrix.



Unlike in previous chapters the least squares estimators used here are not regularized. This eases the calculations and the lack of regularization will not harm us in future applications.

For this choice we showed that for any $a \in \mathbb{R}^d$ it holds that

$$\mathbb{P}\left(\langle \hat{\theta} - \theta_*, a \rangle \geq \sqrt{2\|a\|_{V^{-1}}^2 \log\left(\frac{1}{\delta}\right)}\right) \leq \delta. \quad (21.1)$$

For our purposes, both A_t and x will usually be actions from some (possibly infinite) set $\mathcal{A} \subset \mathbb{R}^d$ and the question of interest is finding the shortest sequence of exploratory actions A_1, \dots, A_n such that the confidence bound in the previous display is smaller than some threshold for all $a \in \mathcal{A}$. To solve this exactly is likely an intractable exercise in integer programming. But a highly accurate approximation turns out to be efficient for a broad class of action sets. Let $\pi : \mathcal{A} \rightarrow [0, 1]$ be a distribution on \mathcal{A} so that $\sum_{a \in \mathcal{A}} \pi(a) = 1$ and $V(\pi) \in \mathbb{R}^{d \times d}$ and $g(\pi) \in \mathbb{R}$ be given by

$$V(\pi) = \sum_{a \in \mathcal{A}} \pi(a) a a^\top, \quad g(\pi) = \max_{a \in \mathcal{A}} \|a\|_{V(\pi)^{-1}}^2. \quad (21.2)$$

In the subfield of statistics called optimal experimental design, the distribution π is called a **design** and the problem of finding the π that minimizes g is called the **G-optimal design problem**. So how to use this? Suppose that π is a design

and $a \in \text{Supp}(\pi)$ and

$$n_a = \left\lceil \frac{\pi(a)g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) \right\rceil. \quad (21.3)$$

Then choosing each action $a \in \text{Supp}(\pi)$ exactly n_a times is enough to ensure that

$$V = \sum_{a \in \text{Supp}(\pi)} n_a aa^\top \geq \frac{g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) V(\pi),$$

which by Eq. (21.1) means that for any $a \in \mathcal{A}$, with probability $1 - \delta$,

$$\langle \hat{\theta} - \theta_*, a \rangle \leq \sqrt{\|a\|_{V^{-1}}^2 \log\left(\frac{1}{\delta}\right)} \leq \varepsilon.$$

By Eq. (21.3) the total number of actions required to ensure a confidence width of no more than ε is bounded by

$$n = \sum_{a \in \text{Supp}(\pi)} n_a = \sum_{a \in \text{Supp}(\pi)} \left\lceil \frac{\pi(a)g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) \right\rceil \leq |\text{Supp}(\pi)| + \frac{g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right).$$

So how big are $|\text{Supp}(\pi)|$ and $g(\pi)$? As we will now show, there exists a π^* that minimizes $g(\pi)$ such that $g(\pi) = d$ and $|\text{Supp}(\pi)| \leq d(d+3)/2$. The first of these facts follows from the following theorem, while the latter is explained in Section 21.2.

THEOREM 21.1 (Kiefer–Wolfowitz) *The following are equivalent:*

- 1 π^* is a minimizer of g .
- 2 π^* is a minimizer of $f(\pi) = -\log \det V(\pi)$.
- 3 $g(\pi^*) = d$.

The theorem shows that G -optimal design is equivalent to the **D-optimal design problem** (D for ‘determinant’), which is the objective in item (2) and (as we shall soon see) has a useful geometric interpretation.

21.1 Proof of Kiefer–Wolfowitz (\dagger)

We follow the original proof by [Kiefer and Wolfowitz \[1960\]](#), which is direct and relies only on elementary linear algebra, convexity and calculus. Nevertheless, this section is not core to the rest of the book and could be skipped on a first pass. To begin we note that since no matter how the exploration distribution π is chosen it holds that $\sum_a \pi(a) \|a\|_{V(\pi)^{-1}}^2 = d$. Hence for all π there exists an $a \in \mathcal{A}$ such that $\|a\|_{V(\pi)^{-1}}^2 \geq d$. The proof will follow by showing that if π maximizes $\log \det(V(\pi))$, then $\|a\|_{V(\pi)^{-1}}^2 \leq d$ for all $a \in \mathcal{A}$. Suppose that π is a distribution

such that $\det V(\pi) > 0$ and

$$\frac{\partial}{\partial \alpha} \log \det V((1-\alpha)\pi + \alpha\pi') \Big|_{\alpha=0} \leq 0 \quad \text{for all distributions } \pi'. \quad (21.4)$$

Let us momentarily fix an alternative distribution π' and let A be a matrix such that $AV(\pi)A^\top = I$ and $AV(\pi')A^\top = B$ where B is diagonal with elements b_1, \dots, b_d (such a matrix exists by simultaneous diagonalization). Then

$$\det V((1-\alpha)\pi + \alpha\pi') = \frac{\prod_{i=1}^d (1 - \alpha + \alpha b_i)}{(\det A)^2}.$$

By noting that the sum of concave functions is concave and checking that $\log(1 - \alpha + \alpha b_i)$ is concave it follows that $\log \det V((1-\alpha)\pi + \alpha\pi')$ is concave in $\alpha \in [0, 1]$. It follows that for π with $\det V(\pi) > 0$ and satisfying Eq. (21.4) that $\pi = \operatorname{argmax}_\pi \log \det(V(\pi))$. The next step is a direct calculation of the derivative in Eq. (21.4) (details see Exercise 21.1):

$$\frac{\partial}{\partial \alpha} \log \det V((1-\alpha)\pi + \alpha\pi') \Big|_{\alpha=0} = \sum_{ij} V(\pi)_{ij}^{-1} V(\pi')_{ij} - d. \quad (21.5)$$

By letting $\pi'(a) = 1$ for some $a \in \mathcal{A}$ we have $\|a\|_{V(\pi)^{-1}}^2 = \sum_{ij} V(\pi)_{ij}^{-1} V(\pi')_{ij} \leq d$.

21.2 Minimum volume ellipsoids and John's theorem (\dagger)

This section depends on a little background on convex optimization and especially the notion of duality. The classic reference is by [Boyd and Vandenberghe \[2004, Chap 5\]](#). Let S_{++}^d be the space of (symmetric) positive definite matrices and recall that a d -dimensional ellipsoid is determined by its center $x_\circ \in \mathbb{R}^d$ and a positive definite matrix $H \in S_{++}^d$ and defined by $E(x_\circ, H) = \{x \in \mathbb{R}^d : \|x - x_\circ\|_{H^{-1}} \leq 1\}$. Given a closed convex set $\mathcal{K} \subset \mathbb{R}^d$ it is a problem in convex geometry to find the ellipsoid E of smallest volume such that $\mathcal{K} \subseteq E$. Such an ellipsoid is called the **minimum-volume enclosing ellipsoid** (MVEE). The volume of an ellipsoid is easily evaluated by noting that if $L = \sqrt{H}$, then $\operatorname{vol}(E(x_\circ, H)) = \operatorname{vol}(E(0, H))$ and $E(0, H) = LB_2^d$ where $B_2^d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ is the unit ball and $LB_2^d = \{Lx : x \in B_2^d\} \subset \mathbb{R}^d$. Therefore

$$\operatorname{vol}(E(x_\circ, H)) = \operatorname{vol}(B_2^d) \det(L) = \operatorname{vol}(B_2^d) \sqrt{\det(H)}.$$

To make the connection to optimal design we consider a modification of the problem of finding the MVEE by adding the restriction that the ellipsoid must be centered ($x_\circ = 0$), which is written as the following convex optimization problem:

$$\begin{aligned} & \min_{H \in S_{++}^d} \log \det(H) \\ & \text{subject to } \mathcal{K} \subseteq E(0, H). \end{aligned}$$

If $\mathcal{K} = \text{co}(\mathcal{A})$ is the convex hull of \mathcal{A} , then the dual of this problem is equivalent to the D -optimal design problem where the Lagrange multipliers play the role of the design π . The dual is

$$\begin{aligned} & \max \log \det \left(\sum_{a \in \mathcal{A}} \lambda(a) aa^\top \right) - \sum_{a \in \mathcal{A}} \lambda(a) + d \\ & \text{subject to } \lambda(a) \geq 0 \text{ for all } a \in \mathcal{A}. \end{aligned} \quad (21.6)$$

As it happens this is one situation where strong duality holds, so the optimization problems are essentially equivalent. By introducing $\pi(a) = \lambda(a) / \sum_{a' \in \mathcal{A}} \lambda(a')$ it is easy to check (again by duality) that the above is equivalent to

$$\begin{aligned} & \max \log \det \left(\sum_{a \in \mathcal{A}} \pi(a) aa^\top \right) + d \log d \\ & \text{subject to } \pi \text{ being a distribution on } \mathcal{A}. \end{aligned}$$

Of course $d \log d$ does not depend on π , so this optimization problem is now equivalent to the D -optimal design problem that appeared in Theorem 21.1. Fritz John's celebrated result concerns the properties of the MVEE with no restriction on the center.

THEOREM 21.2 (John's theorem) *Let $\mathcal{K} \subset \mathbb{R}^d$ be convex, closed and assume that $\text{span}(\mathcal{K}) = \mathbb{R}^d$. Then there exists a unique MVEE of \mathcal{K} . Furthermore, this MVEE is the unit ball B_2^d if and only if there exists $m \leq d(d+3)/2$ contact points ("the core set") u_1, \dots, u_m that belong to both \mathcal{K} and the surface of B_2^d and there also exist positive reals c_1, \dots, c_m such that*

$$\sum_i c_i u_i = 0 \quad \text{and} \quad \sum_i c_i u_i u_i^\top = I, \quad (21.7)$$

To apply John's theorem we first massage the action set so that the MVEE provided by the theorem is centered, but without affecting the optimal design. Let $\mathcal{A}' = \{a : a \in \mathcal{A} \text{ or } -a \in \mathcal{A}\}$ and $\mathcal{K} = \text{co}(\mathcal{A}')$ be the convex hull of \mathcal{A}' . Now take $E = E(x_\circ, H)$ to be the MVEE of \mathcal{K} , which by construction is centered so that $x_\circ = 0$. If $L = \sqrt{H}$, then the image of E under L^{-1} is B_2^d , which is the MVEE of convex set $L^{-1}\mathcal{K}$. Therefore by John's theorem there exists $u_1, \dots, u_m \in L^{-1}\mathcal{K} \cap \partial B_2^d$ and positive reals c_1, \dots, c_m such that Eq. (21.7) holds. In fact, by the curvature of the ellipse we have $u_i \in L^{-1}\mathcal{A}' \cap \partial B_2^d$. Since the trace of a matrix is invariant under rotation,

$$d = \text{trace} \left(\sum_i c_i u_i u_i^\top \right) = \sum_i c_i \text{trace}(u_i u_i^\top) = \sum_i c_i.$$

This allows us to take

$$\pi(a) = \frac{1}{d} \sum_i c_i \mathbb{I}\{Lu_i = a \vee (Lu_i = -a \wedge -a \notin \mathcal{A})\},$$

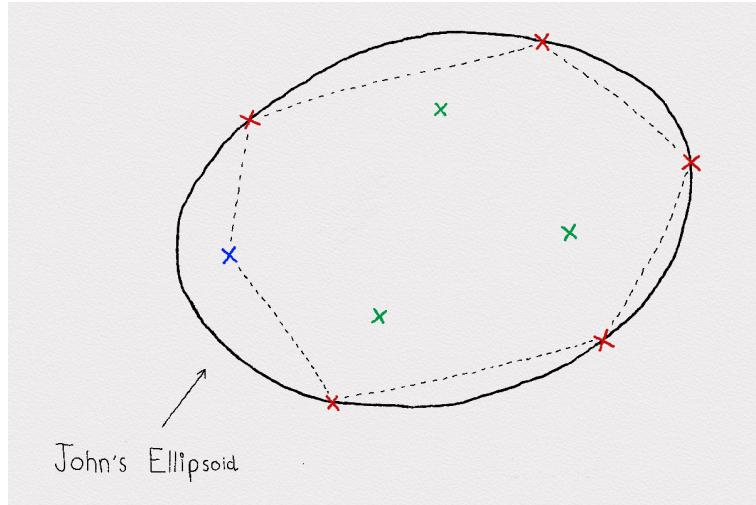


Figure 21.1 John’s ellipsoid for the convex polytope (dashed line) over a small action set. The core set is marked in red. Actions on the boundary of the polytope (and not the core set) are blue, while the green actions are called interior points.

where the complicated expression is due to the fact that a and $-a$ might sometimes both be in \mathcal{A} . Therefore $V(\pi) = LL^\top/d$ and so

$$\sup_{a \in \mathcal{A}} \|a\|_{V(\pi)^{-1}}^2 \leq \sup_{u: \|u\|_2=1} \|L^{-1}(a)\|_{V(\pi)^{-1}}^2 = d \sup_{u: \|u\|_2=1} u^\top L^{-1}LL^\top(L^\top)^{-1}u = d.$$

None of this is terribly surprising in light of Kiefer–Wolfowitz theorem, but John’s theorem also provides a guarantee on the size of the core set, which means that the support of the G -optimal design π can be assumed to have cardinality at most $d(d+3)/2$.

21.3 Notes

- 1 In no applications will we require an exact solution to the design problem. In fact, finding a distribution π such that $g(\pi) \leq (1 + \varepsilon)g(\pi^*)$ will increase the regret of our algorithms by a factor of just $(1 + \varepsilon)^{1/2}$.
- 2 When the action set is finite the computation of the optimal design is a convex problem for which there are numerous efficient approximation algorithms. The Franke-Wolfe algorithm is one such algorithm, which is known as Wynn’s method in optimal experimental design and can be used to find a near-optimal solution for modestly sized problems. More sophisticated methods have also been investigated. A good place to start is: [Vandenberghe et al. \[1998\]](#). If the action set is infinite, then the optimal design can often still be approximated efficiently. The most notable case is when there exists an efficient algorithm (a ‘membership oracle’) for the function $\mathbb{I}\{x \in \mathcal{K}\}$ for any $x \in \mathbb{R}^d$. For details on

this (and many other interesting algorithms involving convexity) see the book by [Grötschel et al. \[2012\]](#).

- 3 While the proof of Theorem 21.1 is sufficiently elementary to include here, we do not know of a simple proof of John’s theorem. Perhaps a reason for the additional difficulty is that John’s proof implicitly shows that the cardinality of the core set is at most $d(d + 3)/2$, which is not revealed at all by Kiefer–Wolfowitz. A proof of John’s theorem may be found in the short book by [Ball \[1997\]](#).

21.4 Bibliographic remarks

According to our best knowledge, the connection to optimal experimental design through the Kiefer–Wolfowitz theorem and the proof that solely relied on this result has not been pointed out in the literature beforehand, though the connection between the Kiefer–Wolfowitz theorem and MVEEs is well known. Besides the previously mentioned book by [Boyd and Vandenberghe \[2004\]](#) there is also a recent book by [Todd \[2016\]](#) that discusses algorithmic issues as well as the duality. The theorem of Kiefer and Wolfowitz is due to them: [Kiefer and Wolfowitz \[1960\]](#). John’s theorem is due to [John \[1948\]](#). The duality mentioned in the text was proved by [Silvey and Sibson \[1972\]](#).

21.5 Exercises

21.1 Prove the correctness of the derivative in Eq. (21.5).

-  Use the fact that the inverse of matrix A is $A^{-1} = M / \det(A)$ where M is the matrix of cofactors of A .

21.2 Find John’s ellipsoid for each of the following sets and use it to derive the G -optimal design.

- (a) The simplex: $\mathcal{K} = \text{co}\{e_1, \dots, e_d\}$.
- (b) The hypercube: $\mathcal{K} = \{x : \|x\|_\infty \leq 1\}$.

21.3 Write a program that accepts as parameters a finite set $\mathcal{A} \subset \mathbb{R}^d$ and returns the G -optimal design $\pi : \mathcal{A} \rightarrow [0, 1]$ that minimizes $g(\pi)$ given in Eq. (21.2).

-  The easiest ‘pure’ way to do this is to implement the Franke-Wolfe algorithm (see the notes). For more robust results we suggest a convex optimization library be used.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

22 Stochastic Linear Bandits with Finitely Many Arms

The optimal design problem has immediate applications to stochastic linear bandits. In Chapter 19 we developed a linear version of the upper confidence bound algorithm that achieves a regret of $R_n = \tilde{O}(d\sqrt{n})$. The only required assumptions were that the sequence of available action-sets were bounded. In this short chapter we consider a more restricted setting where:

- 1 *Fixed finite action set*: The set of actions available in round t is $\mathcal{A} \subset \mathbb{R}^d$ and $|\mathcal{A}| = K$ for some natural number K .
- 2 *Subgaussian rewards*: The reward is $X_t = \langle \theta_*, A_t \rangle + \eta_t$ where η_t is conditionally 1-subgaussian:

$$\mathbb{E}[\exp(\lambda\eta_t)|A_1, \eta_1, \dots, A_{t-1}] \leq \exp(\lambda^2/2) \quad \text{almost surely for all } \lambda \in \mathbb{R}.$$

- 3 *Bounded mean rewards*: $\Delta_a = \max_{b \in \mathcal{A}} \langle \theta_*, b - a \rangle \leq 1$ for all $a \in \mathcal{A}$.

The key difference is that now the set of actions is finite and does not change with time. Under these conditions it becomes possible to design a policy such that

$$R_n = O\left(\sqrt{dn \log(nK)}\right).$$

When K is small this bound improves the regret by a factor of $d^{1/2}$, which in some regimes is large enough to be worth the effort. The core idea is to introduce phases of determinism into the algorithm so that within each phase the actions are chosen independently from the rewards. This decoupling allows us to make use of the tighter confidence bounds available in the fixed design setting as discussed in the previous chapter. The choice of policy within each phase uses the solution to an optimal design problem to minimize the number of required samples to eliminate arms that are far from optimal.

THEOREM 22.1 *With probability at least $1 - \delta$ the regret of Algorithm 11 is at most:*

$$R_n \leq C \sqrt{nd \log\left(\frac{K \log(n)}{\delta}\right)},$$

where $C > 0$ is a universal constant. If $\delta = O(1/n)$, then $\mathbb{E}[R_n] \leq C \sqrt{nd \log(Kn)}$ for appropriately chosen universal constant $C > 0$.

Input $\mathcal{A} \subset \mathbb{R}^d$ and δ

Step 0 Set $\ell = 1$ and let $\mathcal{A}_1 = \mathcal{A}$

Step 1 Let $t_\ell = t$ be the current timestep and find G -optimal design $\pi_\ell : \mathcal{A}_\ell \rightarrow [0, 1]$ that maximizes

$$\log \det V(\pi_\ell) \text{ subject to } \sum_{a \in \mathcal{A}_\ell} \pi_\ell(a) = 1$$

Step 2 Let $\varepsilon_\ell = 2^{-\ell}$ and

$$T_\ell(a) = \left\lceil \frac{2\pi(a)}{\varepsilon_\ell^2} \log \left(\frac{K\ell(\ell+1)}{\delta} \right) \right\rceil \text{ and } T_\ell = \sum_{a \in \mathcal{A}_\ell} T_\ell(a)$$

Step 3 Choose each action $a \in \mathcal{A}_\ell$ exactly $T_a(\ell)$ times

Step 4 Calculate empirical estimate:

$$\hat{\theta} = V_\ell^{-1} \sum_{t=t_\ell}^{t_\ell+T_\ell} A_t X_t$$

Step 5 Eliminate low rewarding arms:

$$\mathcal{A}_{\ell+1} = \left\{ a \in \mathcal{A}_\ell : \max_{b \in \mathcal{A}_\ell} \langle \hat{\theta}, b - a \rangle \geq 2\varepsilon_\ell \right\}.$$

Algorithm 11: Phased elimination with G -optimal exploration

The proof of this theorem follows relatively directly from the high-probability correctness of the confidence intervals used to eliminate low-rewarding arms. We leave the details to the reader in Exercise 22.1.

22.1 Bibliographic remarks

Algorithm 11 is a combination of several existing ideas. The use of phases to decouple the dependence of the design and the outcomes is originally due to [Auer \[2002\]](#), where a more complicated version of the presented problem is solved in which the action set is permitted to change with time. The complexity of the analysis unfortunately prohibited us from presenting these ideas here. Phased approaches have since appeared in many places, but the most similar is the work on spectral bandits by [Valko et al. \[2014\]](#). Neither of these works used the Kiefer–Wolfowitz theorem. This idea is taken from the literature on adversarial linear bandits where John’s ellipsoid has been used to define exploration policies [[Bubeck et al., 2012](#)]. For more details on adversarial linear bandits read on to Part VI.

SupLinRel, LinRel, [Chu et al. \[2011\]](#).

22.2 Exercises

22.1 In this exercise you will prove Theorem 22.1.

- (a) The first step is to use Theorem 21.1 (and the preceding comments) to show that the length of the ℓ th phase is bounded by

$$T_\ell \leq \frac{2d}{\varepsilon_\ell^2} \log \left(\frac{K\ell(\ell+1)}{\delta} \right) + \frac{d(d+3)}{2}$$

- (b) Let $a^* \in \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \theta_* \rangle$ be the optimal arm and use Theorem 21.1 to show that

$$\mathbb{P}(\text{exists phase } \ell \text{ such that } a^* \notin A_\ell) \leq \frac{\delta}{K}.$$

- (c) For action a define $\ell_a = \min\{\ell : \Delta_a < 2\varepsilon_\ell\}$ to be the first phase where the suboptimality gap of arm a is smaller than $2\varepsilon_\ell$. Show that

$$\mathbb{P}(a \in \mathcal{A}_{\ell_a}) \leq \frac{\delta}{K}$$

- (d) Show that with probability at least $1 - \delta$ the regret is bounded by

$$R_n \leq C \sqrt{dn \log \left(\frac{K}{\delta} \right)},$$

where $C > 0$ is a universal constant.

- (e) Show that this implies Theorem 22.1 for the given choice of δ .

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

23 Stochastic Linear Bandits with Sparsity

In Chapter 19 we showed the linear variant of UCB has regret bounded by

$$R_n = O(d\sqrt{n \log(n)}),$$

which for fixed finite action sets can be improved to

$$R_n = \tilde{O}(\sqrt{dn \log(nK)}).$$

For moderately sized action sets these approaches lead to a big improvement over what could be obtained by using the policies that do not make use of the linear structure.

The situation is still not perfect though. In typical applications d is the dimension of the feature space in which the actions are embedded. The features are chosen by the users of the system and one can easily imagine the user has many candidate features and little knowledge about which will be most useful. This presents the user with a challenging tradeoff. If they include many features, then the regret bound will be large. But if a useful feature is omitted, then the linear model will almost certainly be quite wrong. Ideally, one should be able to add features without suffering much additional regret if the feature added does not contribute in a significant way. This can be captured by the notion of sparsity, which is the central theme of this chapter.

23.1 Sparse linear stochastic bandits

The sparse linear stochastic bandit problem is the same as the stochastic linear bandit problem with a small difference. Just like in the standard setting, at the beginning of a round with index t the learner receives a decision set $\mathcal{A}_t \subset \mathbb{R}^d$. They then choose an action $A_t \in \mathcal{A}_t$ and receives the reward

$$X_t = \langle A_t, \theta_* \rangle + \eta_t, \tag{23.1}$$

where $(\eta_t)_t$ is zero-mean noise and $\theta_* \in \mathbb{R}^d$ is an unknown vector. The only difference in the sparse setting is that the parameter vector θ is assumed to have many zero entries. Given $\theta \in \mathbb{R}^d$ let

$$\|\theta\|_0 = \sum_{i=1}^d \mathbb{I}\{\theta_i \neq 0\},$$

which is sometimes called the 0-“norm” (quotations because it is not really a norm, see Exercise 23.1). For the remainder of this chapter we will assume that

- 1 (*Sparse parameter*) There exist known constants M_0 and M_2 such that $\|\theta_*\|_0 \leq M_0$ and $\|\theta_*\|_2 \leq M_2$.
- 2 (*Bounded mean rewards*): $\langle a, \theta_* \rangle \leq 1$ for all $a \in \mathcal{A}_t$ and all rounds t .
- 3 (*Subgaussian noise*): The reward is $X_t = \langle A_t, \theta_* \rangle + \eta_t$ where $\eta_t | \mathcal{F}_{t-1} \sim \text{subG}(1)$ for $\mathcal{F}_t = \sigma(A_1, \eta_1, \dots, A_t, \eta_t)$.

Much ink has been spilled on what can be said about the speed of learning in linear models like (23.1) when $(A_t)_t$ are passively generated and the parameter vector is known to be sparse. Most results are phrased about recovering θ_* , but there also exist a few results that quantify the speed at which good predictions can be made. The ideal outcome would be that the learning speed depends mostly on M_0 , while the dependence on d becomes less severe. Almost all the results come under the assumption that the Grammian of the actions $(A_t)_t$ is well-conditioned.



The **condition number** of a positive definite matrix A is the ratio of its largest and smallest eigenvalues. A matrix is **well conditioned** if it has a small condition number.

The details are a bit more complicated than just the conditioning, but the main point is that the usual assumptions imposed on the Grammian for passive learning are never satisfied when the actions are chosen by a good bandit policy. The reason is simple. Bandit algorithms want to choose the optimal action as often as possible, which means the Grammian will have an eigenvector that points (approximately) towards the optimal action with a large corresponding eigenvalue. We need some approach that does not rely on such strong assumptions.

23.2 Elimination on the hypercube

As a warmup problem we consider the special case where the action set is the d -dimensional hypercube: $\mathcal{A} = [-1, 1]^d$. To reduce clutter we will denote the true parameter vector by θ . As usual, in each round t the learner chooses $A_t \in \mathcal{A}$ and receives reward $X_t = \langle A_t, \theta \rangle + \eta_t$. We make the following standard assumptions:

- 1 (*Bounded mean rewards*): $\|\theta\|_1 \leq 1$, which ensures that $|\langle a, \theta \rangle| \leq 1$ for all $a \in \mathcal{A}$.
- 2 (*Subgaussian noise*): η_t is conditionally 1-subgaussian given the past:

$$\mathbb{E} [\exp(\lambda \eta_t) | \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2}{2}\right) \quad \text{almost surely for all } \lambda \in \mathbb{R},$$

where $\mathcal{F}_{t-1} = \sigma(A_1, X_1, \dots, A_{t-1}, X_{t-1}, A_t)$.



Since conditional subgaussianity comes up frequently, we introduce a notation for it. When X is σ -subgaussian given some σ -field \mathcal{F} we will write $X|\mathcal{F} \sim \text{subG}(\sigma)$. Our earlier statement that the sum of independent subgaussian random variables is subgaussian with a subgaussianity factor that is the sum of the two factors also holds for conditionally subgaussian random variables.

The hypercube is notable as an action set because it enjoys perfect separability. For each dimension $i \in [d]$ the value of $A_{ti} \in [-1, 1]$ can be chosen without regard to the choice of A_{tj} for other dimensions j . The first consequence of this is that the optimal action is $a^* = \text{sign}(\theta)$ where

$$\text{sign}(\theta)_i = \text{sign}(\theta_i) = \begin{cases} 1 & \text{if } \theta_i > 0 \\ 0 & \text{if } \theta_i = 0 \\ -1 & \text{if } \theta_i < 0. \end{cases}$$

So learning the optimal action amounts to learning the sign of θ_i for each dimension. A disadvantage of this structure is that in the worst case the sign of each θ_i must be learned independently, which in Chapter 24 we show leads to a worst case regret of $R_n = \Omega(d\sqrt{n})$. On the positive side, the separability means that θ_i can be estimated in each dimension independently while paying absolutely no price for this experimentation when $\theta_i = 0$. It turns out that this allows us to design a policy whose regret scales with $O(\|\theta\|_0\sqrt{n})$ even without knowing the value of $\|\theta\|_0$.

Let $\mathcal{G}_t = \sigma(A_1, X_1, \dots, A_{t-1}, X_{t-1})$ be the σ -algebra containing information up to time $t-1$ (this differs from \mathcal{F}_t , which also includes information about the action chosen). Now suppose that $(A_{ti})_i$ are chosen to be conditionally independent given \mathcal{G}_t and further assume for some specific $i \in [d]$ that A_{ti} is sampled from a Rademacher distribution so that $\mathbb{P}(A_{ti} = 1|\mathcal{G}_t) = \mathbb{P}(A_{ti} = -1) = 1/2$. Then

$$\begin{aligned} \mathbb{E}[A_{ti}X_t|\mathcal{G}_t] &= \mathbb{E}\left[A_{ti}\left(\sum_{j=1}^d A_{tj}\theta_j + \eta_t\right)\right] \\ &= \theta_i\mathbb{E}[A_{ti}^2|\mathcal{G}_t] + \sum_{j \neq i} \theta_j\mathbb{E}[A_{tj}A_{ti}|\mathcal{G}_t] + \mathbb{E}[\eta_t|\mathcal{G}_t] = \theta_i, \end{aligned}$$

where the first equality is the definition of $X_t = \langle A_t, \theta \rangle + \eta_t$, the second by linearity of expectation and the third by the conditional independence of $(A_{ti})_i$ and the fact that $\mathbb{E}[A_{ti}|\mathcal{G}_t] = 0$ and $\mathbb{E}[A_{ti}^2|\mathcal{G}_t] = 1$. This looks quite promising, but we should also check the variance. Using our assumptions we have: $\mathbb{E}[\eta] = 0$ and $\mathbb{E}[\eta^2] \leq 1$ and $\langle a, \theta \rangle \leq 1$ for all actions a we have

$$\mathbb{V}[A_{ti}X_t|\mathcal{G}_t] = \mathbb{E}[A_{ti}^2X_t^2|\mathcal{G}_t] - \theta_i^2 = \mathbb{E}[(\langle A_t, \theta \rangle + \eta)^2|\mathcal{G}_t] - \theta_i^2 \leq 2. \quad (23.2)$$

And now we have cause for celebration. The value of θ_i can be estimated by choosing A_{ti} to be a Rademacher random variable independent of the choices in

other dimensions. All the policy does is treat all dimensions independently. For a particular dimension (say i) it explores by choosing $A_{ti} \in \{-1, 1\}$ uniformly at random until its estimate is sufficiently accurate to commit to either $A_{ti} = 1$ or $A_{ti} = -1$ for all future rounds. How long this takes depends on $|\theta_i|$, but note that if $|\theta_i|$ is small, then the price of exploring is also limited. The policy that results from this idea is called Selective Explore-Then-Commit (Algorithm 12, SETC).

```

1: Input  $n$  and  $d$ 
2: Set  $E_{1,i} = 1$  and  $\mathcal{C}_{1,i} = \mathbb{R}$  for all  $i \in [d]$ 
3: for  $t = 1, \dots, n$  do
4:   For each  $i \in [d]$  sample  $B_{ti} \sim \text{RADEMACHER}$ 
5:   Choose action:

$$(\forall i) \quad A_{ti} = \begin{cases} B_{ti} & \text{if } 0 \in \mathcal{C}_{ti} \\ 1 & \text{if } \mathcal{C}_{ti} \subset (0, \infty] \\ -1 & \text{if } \mathcal{C}_{ti} \subset [-\infty, 0). \end{cases}$$

6:   Play  $A_t$  and observe  $X_t$ 
7:   Construct empirical estimators:

$$(\forall i) \quad T_i(t) = \sum_{s=1}^t E_{si} \quad \hat{\theta}_{ti} = \frac{\sum_{s=1}^t E_{si} A_{si} X_s}{T_i(t)}$$

8:   Construct confidence intervals:

$$(\forall i) \quad W_{ti} = 2 \sqrt{\left( \frac{1}{T_i(t)} + \frac{1}{T_i(t)^2} \right) \log(n \sqrt{2T_i(t) + 1})}$$


$$(\forall i) \quad \mathcal{C}_{t+1,i} = [\hat{\theta}_{ti} - W_{ti}, \hat{\theta}_{ti} + W_{ti}]$$

9:   Update exploration parameters:

$$(\forall i) \quad E_{t+1,i} = \begin{cases} 0 & \text{if } 0 \notin \mathcal{C}_{t+1,i} \text{ or } E_{ti} = 0 \\ 1 & \text{otherwise.} \end{cases}$$

10: end for
```

Algorithm 12: Selective Explore-Then-Commit

THEOREM 23.1 *There exists a universal constant $C > 0$ such that the regret of SETC satisfies:*

$$R_n \leq 2\|\theta\|_1 + C \sum_{i:\theta_i \neq 0} \frac{\log(n)}{|\theta_i|}.$$

Furthermore $R_n \leq C\|\theta\|_0 \sqrt{n \log(n)}$.

By appealing to the central limit theorem and the variance calculation in

Eq. (23.2) we should be hopeful that the confidence intervals used by the algorithm are sufficiently large to contain the true θ_i with high probability, but this still needs to be proven.

LEMMA 23.1 *Define $\tau_i = n \wedge \max\{t : E_{ti} = 1\}$ and $F_i = \mathbb{I}\{\tau_i \leq n \wedge \theta_i \notin \mathcal{C}_{\tau_i+1,i}\}$ be the event that θ_i is not in the confidence interval constructed at time τ_i . Then $\mathbb{P}(F_i) \leq 1/n$.*

Leaving the proof of Lemma 23.1 to the next section, we first use it to prove Theorem 23.1.

Proof of Theorem 23.1 Recalling the definition of the regret and using the fact that the optimal action is $a^* = \text{sign}(\theta)$ we have the following regret decomposition.

$$R_n = \max_{a \in \mathcal{A}} \langle a, \theta \rangle - \mathbb{E} \left[\sum_{t=1}^n \langle A_t, \theta \rangle \right] = \sum_{i=1}^d \underbrace{\left(n|\theta_i| - \mathbb{E} \left[\sum_{t=1}^n A_{ti} \theta_i \right] \right)}_{R_{ni}}. \quad (23.3)$$

Clearly if $\theta_i = 0$, then $R_{ni} = 0$. And so it suffices to bound R_{ni} for each i with $|\theta_i| > 0$. Suppose that $|\theta_i| > 0$ for some i and the failure event F_i given in Lemma 23.1 does not occur. Then $\theta_i \in \mathcal{C}_{\tau_i+1,i}$ and by definition of the algorithm $A_{ti} = \text{sign}(\theta_i)$ for all $t \geq \tau_i$. Therefore

$$\begin{aligned} R_{ni} &= n|\theta_i| - \mathbb{E} \left[\sum_{t=1}^n A_{ti} \theta_i \right] = \mathbb{E} \left[\sum_{t=1}^n |\theta_i| (1 - A_{ti} \text{sign}(\theta_i)) \right] \\ &\leq n|\theta_i| \mathbb{P}(F_i) + |\theta_i| \mathbb{E} [\mathbb{I}\{F_i^c\} \tau_i] \end{aligned} \quad (23.4)$$

Since τ_i is the last round t when $0 \notin \mathcal{C}_{t+1,i}$ it follows that if F_i does not occur, then $\theta \in \mathcal{C}_{\tau_i,i}$ and $0 \in \mathcal{C}_{\tau_i,i}$. Thus the width of the confidence interval $\mathcal{C}_{\tau_i,i}$ must be at least $|\theta_i|$ and so

$$2W_{\tau_i-1} = 4 \sqrt{\left(\frac{1}{\tau_i-1} + \frac{1}{(\tau_i-1)^2} \right) \log(n\sqrt{2\tau_i-1})} \geq |\theta_i|,$$

which after rearranging shows for some universal constant $C > 0$ that

$$\mathbb{I}\{F_i^c\} (\tau_i - 1) \leq 1 + \frac{C \log(n)}{|\theta_i|^2}.$$

Combining this result with Eq. (23.4) leads to

$$R_{ni} \leq n|\theta_i| \mathbb{P}(F_i) + |\theta_i| + \frac{C \log(n)}{|\theta_i|}.$$

Using Lemma 23.1 to bound $\mathbb{P}(F_i)$ and substituting into the decomposition Eq. (23.3) completes the proof of the first part. The second part is left as an exercise to the reader. \square

23.3 Proof of technical lemma

We start with a simple variation on the self-normalized concentration inequality of Theorem 20.1.

LEMMA 23.2 *Let $\delta \in (0, 1)$ and $(\mathcal{F}_t)_{t \in [n]}$ be a filtration and $(Z_t)_{t \in [n]}$ be \mathcal{F}_t -adapted such that $Z_t | \mathcal{F}_{t-1} \sim \text{subG}(\sigma)$. Then for any stopping time $\tau \in [n]$ it holds that*

$$\mathbb{P} \left(\text{exists } t \leq \tau : |S_t| \geq \sqrt{2\sigma^2(t+1) \log \left(\frac{\sqrt{t\sigma^2 + 1}}{\delta} \right)} \right) \leq \delta.$$

Proof Let $f(\lambda) = \frac{1}{\sqrt{2\pi}} \exp(-\lambda^2/2)$ be the density of the standard Gaussian and define supermartingale M_t by

$$M_t = \int_{\mathbb{R}} f(\lambda) \exp \left(\lambda S_t - \frac{t\sigma^2 \lambda^2}{2} \right) d\lambda = \frac{1}{\sqrt{t\sigma^2 + 1}} \exp \left(\frac{S_t^2}{2\sigma^2(t+1)} \right).$$

Since $\mathbb{E}[M_\tau] = M_0 = 1$, by the maximal inequality $\mathbb{P}(\sup_{t \leq \tau} M_t \geq 1/\delta) \leq \delta$. Rearranging yields the claim.

$$\mathbb{P} \left(\text{exists } t \leq \tau : |S_t| \geq \sqrt{2\sigma^2(t+1) \log \left(\frac{\sqrt{t\sigma^2 + 1}}{\delta} \right)} \right) \leq \delta. \quad \square$$



One might question whether or not the choice of Gaussian for mixing distribution f is optimal. In fact it is nothing more than a convenient choice that allows for an easy evaluation of the integral. By selecting a more appropriate mixing distribution one can show a result that is reminiscent of the law of the iterated logarithm. For details see Exercise 23.6.

Proof of Lemma 23.1 Let $Z_{ti} = A_{ti}\eta_i + A_{ti} \sum_{j \neq i} A_{tj}\theta_j$. Setting $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$, we see that Z_{ti} is \mathcal{F}_t -adapted. Letting $S_{ti} = \sum_{j \neq i} A_{tj}\theta_j$ and expanding $Z_{ti} = A_{ti}S_{ti} + A_{ti}\eta_t$. The first step is to check that $Z_{ti} | \mathcal{F}_{t-1} \sim \text{subG}(\sqrt{2})$.

$$\begin{aligned} \mathbb{E}[\exp(\lambda Z_{ti}) | \mathcal{F}_{t-1}] &= \mathbb{E}[\mathbb{E}[\exp(\lambda Z_{ti}) | \mathcal{F}_{t-1}, A_t] | \mathcal{F}_{t-1}] \\ &= \mathbb{E}[\exp(\lambda A_{ti}S_{ti}) \mathbb{E}[\exp(\lambda A_{ti}\eta_t) | \mathcal{F}_{t-1}, A_t] | \mathcal{F}_{t-1}] \\ &\leq \mathbb{E} \left[\exp(\lambda A_{ti}S_{ti}) \exp \left(\frac{\lambda^2}{2} \right) \middle| \mathcal{F}_{t-1} \right] \\ &= \exp \left(\frac{\lambda^2}{2} \right) \mathbb{E}[\mathbb{E}[\exp(\lambda A_{ti}S_{ti}) | \mathcal{F}_{t-1}, S_{ti}] | \mathcal{F}_{t-1}] \\ &\leq \exp \left(\frac{\lambda^2}{2} \right) \mathbb{E} \left[\exp \left(\frac{\lambda^2 S_{ti}^2}{2} \right) \middle| \mathcal{F}_{t-1} \right] \\ &\leq \exp(\lambda^2), \end{aligned}$$

where the first inequality used that η_t and A_t are conditionally independent given \mathcal{F}_{t-1} and that $\eta_t|\mathcal{F}_{t-1} \sim \text{subG}(1)$, the second to last inequality used that S_{ti} and A_{ti} are conditionally independent given \mathcal{F}_{t-1} and that $A_{ti}|\mathcal{F}_{t-1} \sim \text{subG}(1)$, the last step used that $|S_{ti}| \leq 1$. From this we conclude that $Z_{ti}|\mathcal{F}_{t-1} \sim \text{subG}(\sqrt{2})$. The result follows by applying Lemma 23.2 with stopping time $\tau_i = n \wedge \max\{t : E_{ti} = 1\}$. Then $\mathbb{P}(F_i) = \mathbb{P}(\theta_i \notin \mathcal{C}_{\tau_i, i}) \leq 1/n$. \square

23.4 UCB with sparsity

A new plan is needed to relax the assumption that the action set is a hypercube. The idea is to modify the ellipsoidal confidence set used in Chapter 19 to have a smaller radius, which is made possible by exploiting the lower variance of the least-squares estimator when the unknown parameter is sparse. We will see that modifying the algorithm in Chapter 19 to use the smaller confidence intervals improves the regret to $R_n = O(\sqrt{dpn} \log(n))$.



Without assumptions on the action-set one cannot hope to have a regret smaller than $O(\sqrt{dn})$. To see this, recall that d -armed bandits can be represented as linear bandits with $\mathcal{A}_t = \{e_1, \dots, e_d\}$. For these problems Theorem 15.1 shows that for any policy there exists a d -armed bandit for which $R_n = \Omega(\sqrt{dn})$. Checking the proof reveals that when adapted to the linear setting the parameter vector is 2-sparse.

23.5 Online to confidence set conversion

The construction that follows makes use of a kind of duality between online prediction and confidence sets. While we will only apply the idea to the sparse linear case, the approach is generic. Unless otherwise mentioned, for the remainder of the chapter we make the following assumptions:

The prediction problem considered is **online linear prediction** where the prediction error is measured by the squared loss. This is also known as **online linear regression**. The learner interacts with an environment in a sequential manner where in each round $t \in \mathbb{N}^+$:

- 1 The environment chooses $X_t \in \mathbb{R}$ and $A_t \in \mathbb{R}^d$ in an arbitrary fashion.
- 2 The value of A_t is revealed to the learner (but not X_t).
- 3 The learner produces a real-valued prediction \hat{X}_t in some way.
- 4 The environment reveals X_t to the learner and the loss is $(X_t - \hat{X}_t)^2$.

The learner's goal is to produce predictions whose total loss is not much worse

than the loss suffered by any of the linear predictors in some set $\Theta \subset \mathbb{R}^d$. The regret of the learner relative to a linear predictor that uses the weights $\theta \in \mathbb{R}^d$ is

$$\rho_n(\theta) = \sum_{t=1}^n (X_t - \hat{X}_t)^2 - \sum_{t=1}^n (X_t - \langle A_t, \theta \rangle)^2. \quad (23.5)$$

We say that the learner enjoys a regret guarantee B_n relative to Θ if for any strategy of the environment,

$$\sup_{\theta \in \Theta} \rho_n(\theta) \leq B_n. \quad (23.6)$$

The online learning literature in machine learning has a number of powerful algorithms for this learning problem with equally powerful regret guarantees. Later we will give a specific result for the sparse case when $\Theta = \{x : \|x\|_0 \leq M_0\}$, but first we show how to use such a learning algorithm to construct a confidence set. Take any learner for online linear regression and assume the environment generates X_t in a stochastic manner like in linear bandits:

$$X_t = \langle A_t, \theta_* \rangle + \eta_t, \quad (23.7)$$

Combining Eqs. (23.5) to (23.7) with elementary algebra,

$$\begin{aligned} Q_t &= \sum_{t=1}^n (\hat{X}_t - \langle A_t, \theta_* \rangle)^2 = \rho_n(\theta_*) + 2 \sum_{t=1}^n \eta_t (\hat{X}_t - \langle A_t, \theta_* \rangle) \\ &\leq B_n + 2 \sum_{t=1}^n \eta_t (\hat{X}_t - \langle A_t, \theta_* \rangle), \end{aligned} \quad (23.8)$$

where the first equality serves as the definition of Q_t . Let us now take stock for a moment. If we could somehow remove the dependence on the noise η_t in the right hand side, then we could define a confidence set consisting of all θ that satisfy the equation. Of course the noise has zero mean and is conditionally independent of its multiplier, so the expectation of this term is zero. If we can control the fluctuations with high probability, then we will have made some progress. Let

$$Z_t = \sum_{s=1}^t \eta_s (\hat{X}_s - \langle A_s, \theta_* \rangle)$$

Since \hat{X}_t is chosen based on information available at the beginning of the round, \hat{X}_t is \mathcal{F}_{t-1} -measurable and so

$$(Z_t - Z_{t-1}) | \mathcal{F}_{t-1} \sim \text{subG}(\sigma_t), \quad \text{where } \sigma_t^2 = (\hat{X}_t - \langle A_t, \theta_* \rangle)^2.$$

The uniform self-normalized tail bound (Theorem 20.1) with $\lambda = 1$ implies that,

$$\mathbb{P} \left(\text{exists } t \geq 0 \text{ such that } |Z_t| \geq \sqrt{(1+Q_t) \log \left(\frac{1+Q_t}{\delta^2} \right)} \right) \leq \delta.$$

Provided this low probability event does not occur, then from Eq. (23.8) we have

$$Q_t \leq B_t + 2\sqrt{(1+Q_t)\log\left(\frac{1+Q_t}{\delta^2}\right)}. \quad (23.9)$$

While both sides depend on Q_t , the left hand side grows linearly, while the right hand side grows sublinearly in Q_t . This means that the largest value of Q_t that satisfies the above inequality is finite. A tedious calculation then shows this value must be less than

$$\beta_t(\delta) = 1 + 2B_t + 32\log\left(\frac{\sqrt{8} + \sqrt{1+B_t}}{\delta}\right). \quad (23.10)$$

By piecing together the parts we conclude that with probability at least $1 - \delta$ the following holds for all t :

$$Q_t = \sum_{s=1}^t (\hat{X}_s - \langle A_s, \theta_* \rangle)^2 \leq \beta_t(\delta).$$

We could define \mathcal{C}_{t+1} to be the set of all θ such that the above holds with θ_* replaced by θ , but there is one additionally subtlety, which is that the resulting confidence interval may be unbounded (think about the case that $\sum_{s=1}^t A_s A_s^\top$ is not invertible). In Chapter 19 we overcame this problem by regularizing the least squares estimator. Since we have assumed that $\|\theta_*\|_2 \leq M_2$ the previous display implies that

$$\|\theta_*\|_2^2 + \sum_{s=1}^t (\hat{X}_s - \langle A_s, \theta_* \rangle)^2 \leq M_2^2 + \beta_t(\delta).$$

All together we have the following theorem.

THEOREM 23.2 *Let $\delta \in (0, 1)$ and assume that $\theta_* \in \Theta$ and $\sup_{\theta \in \Theta} \rho_t(\theta) \leq B_t$. If*

$$\mathcal{C}_{t+1} = \left\{ \theta \in \mathbb{R}^d : \|\theta\|_2^2 + \sum_{s=1}^t (\hat{X}_s - \langle A_s, \theta \rangle)^2 \leq M_2^2 + \beta_t(\delta) \right\},$$

then $\mathbb{P}(\text{exists } t \in \mathbb{N} \text{ such that } \theta_ \notin \mathcal{C}_{t+1}) \leq \delta$.*

The confidence set in Theorem 23.2 is not in the most convenient form. By defining $V_t = I + \sum_{s=1}^t A_s A_s^\top$ and $S_t = \sum_{s=1}^t A_s \hat{X}_s$ and $\hat{\theta}_t = V_t^{-1} S_t$ and performing an algebraic calculation that we leave to the reader (see Exercise 23.5) one can see that

$$\|\theta\|_2^2 + \sum_{s=1}^t (\hat{X}_s - \langle A_s, \theta \rangle)^2 = \|\theta - \hat{\theta}_t\|_{V_t}^2 + \sum_{s=1}^t (\hat{X}_s - \langle \hat{\theta}_t, A_s \rangle)^2 + \|\hat{\theta}_t\|_2^2. \quad (23.11)$$

Using this, the confidence set can be rewritten in the familiar form of an ellipsoid:

$$\mathcal{C}_{t+1} = \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{V_t}^2 \leq M_2^2 + \beta_t(\delta) - \|\hat{\theta}_t\|_2^2 - \sum_{s=1}^t (\hat{X}_s^2 - \langle \hat{\theta}_t, A_s \rangle)^2 \right\}.$$

```

1: Input Online linear predictor and regret bound  $B_t$ , confidence parameter
    $\delta \in (0, 1)$ 
2: for  $t = 1, \dots, n$  do
3:   Receive action set  $\mathcal{A}_t$ 
4:   Computer confidence set:

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^d : \|\theta\|_2^2 + \sum_{s=1}^{t-1} (\hat{X}_s - \langle A_s, \theta \rangle)^2 \leq M_2^2 + \beta_t(\delta) \right\}$$

5:   Calculate optimistic action

$$A_t = \operatorname{argmax}_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_t} \langle a, \theta \rangle$$

6:   Feed  $A_t$  to the online linear predictor and obtain prediction  $\hat{X}_t$ 
7:   Play  $A_t$  and receive reward  $X_t$ 
8:   Feed  $X_t$  to online linear predictor as feedback
9: end for

```

Algorithm 13: Online Linear Predictor UCB

It is not obvious that \mathcal{C}_{t+1} is not empty because the radius could be negative. Theorem 23.2 shows, however, that with high probability $\theta_* \in \mathcal{C}_{t+1}$. At last we have established all the conditions required for Theorem 19.1, which implies the following theorem bounding the regret of Algorithm 13.

THEOREM 23.3 *With probability at least $1 - \delta$ the pseudo-regret of OLR-UCB satisfies*

$$\hat{R}_n \leq \sqrt{8dn(M_2^2 + \beta_{n-1}(\delta)) \log\left(1 + \frac{n}{d}\right)}.$$

23.6 Sparse online linear prediction

THEOREM 23.4 *There exists a strategy π for the learner such that for any $\theta \in \mathbb{R}^d$, the regret $\rho_n(\theta)$ of π against any strategic environment such that $\max_{t \in [n]} \|A_t\|_2 \leq L$ and $\max_{t \in [n]} |X_t| \leq X$ satisfies*

$$\rho_n(\theta) \leq cX^2\|\theta\|_0 \left\{ \log(e + n^{1/2}L) + C_n \log(1 + \frac{\|\theta\|_1}{\|\theta\|_0}) \right\} + (1 + X^2)C_n,$$

where $c > 0$ is some universal constant and $C_n = 2 + \log_2 \log(e + n^{1/2}L)$.

The strategy is a variant of the exponential weights method except that the method is now adjusted so that the set of experts is now \mathbb{R}^d . An appropriate sparsity prior is used and when predicting an appropriate truncation strategy is used. The details of the procedure are less important at this stage for our purposes and are thus left out. A reference to the work containing the missing details will be given at the end of the chapter.

Note that $A_n = O(\log \log(n))$. Hence, dropping the dependence on X and L , for $p > 0$, $\sup_{\theta: \|\theta\|_0 \leq p, \|\theta\|_2 \leq L} \rho_n(\theta) = O(p \log(n))$. Note how strong this is: The guarantee hold no matter what strategy the environment uses!

Now, in sparse linear bandits with subgaussian noise, the noise $(\eta_t)_t$ is not necessarily bounded, and as a consequence the rewards $(X_t)_t$ are also not necessarily bounded. However, the subgaussian property implies with probability $1 - \delta$, $|\eta_t| \leq \log(2/\delta)$. Now, choosing $\delta = 1/n^2$, we thus see that for problems with bounded mean reward, $\max_{t \in [n]} |X_t| \leq X \doteq 1 + \log(2n^2)$ with probability at least $1 - 1/n$. Putting things together then yields the announced result. The expected regret of OLR-UCB when using the strategy π from above satisfies

$$R_n = \tilde{O}(\sqrt{dpn}).$$

23.7 Notes

- 1 The strategy achieving the bound in Theorem 23.4 is not computationally efficient. In fact we do not know of any polynomial time algorithm with logarithmic regret for this problem. The consequence is that Algorithm 13 does not yet have an efficient implementation.
- 2 While we focused on the sparse case, the results and techniques apply to other settings. For example, we can also get alternative confidence sets from results in online learning even for the standard non-sparse case. Or one may consider additional or different structural assumptions on θ (for example, θ that when reshaped into a matrix, could have a low spectral norm).
- 3 When the online linear regression results are applied it is important to use the tightest possible, data-dependent regret bounds B_n . In online learning most regret bounds start as tight, data-dependent bounds, which are then loosened to get further insight into the structure of problems. For our application, naturally one should use the tightest available regret bounds (or one should attempt to modify the existing proofs to get tighter data-dependent bounds). The gains from using data-dependent bounds can be significant.
- 4 We need to emphasize that the sparsity parameter p must be known in advance and that no algorithm can simultaneously enjoy a regret of $\Omega(\sqrt{dpn})$ for all p simultaneously. This will be seen shortly in Chapter 24 that focuses exclusively on lower bounds for stochastic linear bandits.

23.8 Bibliographical Remarks

The Selective Explore-Then-Commit algorithm is due to the authors [Lattimore et al., 2015]. The construction for the sparse case is from another paper co-authored by one of the authors [Abbasi-Yadkori et al., 2012]. The online linear predictor that competes with sparse parameter vectors and its analysis

summarized in Theorem 23.4 is due to [Gerchinovitz, 2013, Thm. 10]. A recent paper by Rakhlin and Sridharan [2017] also discusses relationship between online learning regret bounds and self-normalized tail bounds of the type given here. Interestingly, what they show is that the relationship goes in both directions: Tail inequalities imply regret bounds and regret bounds imply tail inequalities. We are told by Francesco Orabona that techniques similar to used here for constructing confidence bounds have been used earlier in a series of papers by Claudio Gentile and friends. For completeness, here is the list for further exploration: Dekel et al. [2010, 2012], Crammer and Gentile [2013], Gentile and Orabona [2012, 2014]. Carpentier and Munos [2012] have also published a paper on sparse linear stochastic bandits, but with the action-set restricted to the $(d-1)$ -sphere. Like the hypercube, it turns out that this makes it possible to avoid the poor dependence on the dimension and their regret bound is $R_n = O(p\sqrt{n}\log(d))$. The online-to-confidence set construction idea has recently been used for designing more efficient algorithms for generalized linear bandits [Jun et al., 2017].

23.9 Exercises

23.1 A norm on \mathbb{R}^d is a function $\|\cdot\| : \mathbb{R}^d \rightarrow \mathbb{R}$ such that for all $a \in \mathbb{R}$ and $x, y \in \mathbb{R}^d$ it holds that: (a) $\|x\| = 0$ if and only if $x = 0$ and (b) $\|ax\| = |a|\|x\|$ and (c) $\|x + y\| \leq \|x\| + \|y\|$ and (d) $\|x\| \geq 0$. Show that $\|\cdot\|_0$ given by $\|x\|_0 = \sum_{i=1}^d \mathbb{I}\{x_i \neq 0\}$ is not a norm.

23.2 Prove the second part of Theorem 23.1.



Think about what happens to R_{ni} if $|\theta_i|$ is small.

23.3 Algorithm 12 is not anytime (it requires advance knowledge of the horizon). Design a modified version that does not require this knowledge and prove a comparable regret bound to what was given in Theorem 23.1.



One way is to use the doubling trick, but a more careful approach will lead to a more practical algorithm.

23.4 Complete the calculation to derive Eq. (23.10) from Eq. (23.9).

23.5 Prove the equality in Eq. (23.11).

23.6 Let f be a density function on $[0, \infty)$ so that $\int_0^\infty f(\lambda)d\lambda = 1$ and $f(\lambda) \geq 0$ for all $\lambda \geq 0$. Then define

$$M_n = \int_{\mathbb{R}} f(\lambda) \exp\left(\lambda S_n - \frac{\lambda^2 n}{2}\right) d\lambda.$$

- (a) Show that $\operatorname{argmax}_{\lambda \in \mathbb{R}} \lambda S_n - \lambda^2 n/2 = S_n/n$.
- (b) Suppose that $f(\lambda)$ is monotone decreasing for $\lambda > 0$. Show that for any $\varepsilon > 0$ and $\Lambda_n = S_n/n$ that,

$$M_n \geq \varepsilon \Lambda_n f(\Lambda_n(1 + \varepsilon)) \exp\left(\frac{(1 - \varepsilon^2)S^2}{2n}\right)$$

- (c) Use the previous result to show for any $\delta \in (0, 1)$ that

$$\mathbb{P}\left(\text{exists } n : S_n \geq \inf_{\varepsilon > 0} \sqrt{\frac{2n}{(1 - \varepsilon^2)} \left(\log\left(\frac{1}{\delta}\right) + \log\left(\frac{1}{\varepsilon \Lambda_n f(\Lambda_n(1 + \varepsilon))}\right) \right)}\right) \leq \delta.$$

- (d) Find an f such that $\int_0^\infty f(\lambda)d\lambda = 1$ and $f(\lambda) \geq 0$ for all $\lambda \in \mathbb{R}$ and

$$\log\left(\frac{1}{\lambda f(\lambda)}\right) = (1 + o(1)) \log \log\left(\frac{1}{\lambda}\right)$$

as $\lambda \rightarrow 0$.

- (e) Use the previous results to show that

$$\mathbb{P}\left(\limsup_{n \rightarrow \infty} \frac{S_n}{\sqrt{2n \log \log(n)}} \leq 1\right) = 1.$$

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

24 Minimax Lower Bounds for Stochastic Linear Bandits

Lower bounds for linear bandits turn out to be more nuanced than those for the classical finite-armed bandit. The difference is that for linear bandits the shape of the action set plays a role in the form of the regret, not just the distribution of the noise. This should not come as a big surprise because the stochastic finite-armed bandit problem can be modeled as a linear bandit with actions being the standard basis vectors, $\mathcal{A} = \{e_1, \dots, e_K\}$. In this case the actions are orthogonal, which means that samples from one action do not give information about the rewards for other actions. Other action sets such as the sphere ($\mathcal{A} = S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$) do not share this property. For example, if $d = 2$ and $\mathcal{A} = S^1$ and an algorithm chooses actions $e_1 = (1, 0)$ and $e_2 = (0, 1)$ many times, then it can deduce the reward it would obtain from choosing any other action.

All results of this chapter have a worst-case flavor showing what is (not) achievable in general, or under a sparsity constraint, or if the realizable assumption is not satisfied. The analysis uses the information-theoretic tools introduced in Part IV combined with careful choices of action sets. The hard part is guessing what is the worst case, which is followed by simply turning the crank on the usual machinery.

In all lower bounds we use a simple model with Gaussian noise. For action $A_t \in \mathcal{A} \subseteq \mathbb{R}^d$ the reward is $X_t = \mu(A_t) + \eta_t$ where $\eta_t \sim \mathcal{N}(0, 1)$ is a sequence of independent standard Gaussian noise and $\mu : \mathcal{A} \rightarrow [0, 1]$ is the mean reward. We will usually assume there exists a $\theta \in \mathbb{R}^d$ such that $\mu(a) = \langle a, \theta \rangle$. We write \mathbb{P}_μ to indicate the measure on outcomes induced by the interaction of the fixed policy and the Gaussian bandit parameterised by μ . Because we are now proving lower bounds it becomes necessary to be explicit about the dependence of the regret on \mathcal{A} and μ or θ . The regret of a policy is:

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

where the expectation is taken with respect to \mathbb{P}_μ . Except in Section 24.4 we assume the reward function is linear, which means there exists a $\theta \in \mathbb{R}^d$ such that $\mu(a) = \langle a, \theta \rangle$. In these cases we write $R_n(\mathcal{A}, \theta)$ and \mathbb{E}_θ and \mathbb{P}_θ . Recall the notation used for finite-armed bandits by defining $T_x(t) = \sum_{s=1}^t \mathbb{I}\{A_s = x\}$.

24.1 Hypercube

The first lower bound is for the hypercube action-set and shows that the upper bounds in Chapter 19 cannot be improved in general.

THEOREM 24.1 *Let $\mathcal{A} = [-1, 1]^d$ and $\Theta = \{-n^{-1/2}, n^{-1/2}\}^d$. Then for any policy there exists a $\theta \in \Theta$ such that:*

$$R_n(\mathcal{A}, \theta) \geq \frac{\exp(-2)}{8} d\sqrt{n}.$$

Proof By the relative entropy identity (Lemma 15.1) we have for $\theta, \theta' \in \Theta$ that

$$D(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2} \sum_{t=1}^n \mathbb{E}_\theta [\langle A_t, \theta - \theta' \rangle^2]. \quad (24.1)$$

For $i \in [d]$ and $\theta \in \Theta$ define

$$p_{\theta i} = \mathbb{P}_\theta \left(\sum_{t=1}^n \mathbb{I}\{\text{sign}(A_{ti}) \neq \text{sign}(\theta_i)\} \geq n/2 \right).$$

Now let $i \in [d]$ and $\theta \in \Theta$ be fixed and let $\theta' = \theta$ except for $\theta'_i = -\theta_i$. Then by the high probability version of Pinsker's inequality (Theorem 14.2) and Eq. (24.1),

$$p_{\theta i} + p_{\theta' i} \geq \frac{1}{2} \exp \left(-\frac{1}{2} \sum_{t=1}^n \mathbb{E}_\theta [\langle A_t, \theta - \theta' \rangle^2] \right) \geq \frac{1}{2} \exp(-2). \quad (24.2)$$

Applying an ‘averaging hammer’ over all $\theta \in \Theta$, which satisfies $|\Theta| = 2^d$:

$$\sum_{\theta \in \Theta} \frac{1}{|\Theta|} \sum_{i=1}^d p_{\theta i} = \frac{1}{|\Theta|} \sum_{i=1}^d \sum_{\theta \in \Theta} p_{\theta i} \geq \frac{d}{4} \exp(-2).$$

Since $p_{\theta i}$ is nonnegative this implies there exists a $\theta \in \Theta$ such that $\sum_{i=1}^d p_{\theta i} \geq d \exp(-2)/4$. By the definition of $p_{\theta i}$ the regret for this choice of θ is at least

$$\begin{aligned} R_n(\mathcal{A}, \theta) &\geq \sqrt{\frac{1}{n} \sum_{i=1}^d \mathbb{E} \left[\sum_{t=1}^n \mathbb{I}\{\text{sign}(A_{ti}) \neq \text{sign}(\theta_i)\} \right]} \\ &\geq \frac{\sqrt{n}}{2} \sum_{i=1}^d \mathbb{P}_\theta \left(\sum_{t=1}^n \mathbb{I}\{\text{sign}(A_{ti}) \neq \text{sign}(\theta_i)\} \geq n/2 \right) \\ &= \frac{\sqrt{n}}{2} \sum_{i=1}^d p_{\theta i} \geq \frac{\exp(-2)}{8} d\sqrt{n}. \end{aligned} \quad \square$$



Except for logarithmic factors this shows the algorithm of Chapter 19 is near-optimal for this action set. The same proof works when $\mathcal{A} = \{-1, 1\}^d$ is restricted to the corners of the hypercube, which is a finite-armed linear bandit. In Chapter 22 we gave a policy with regret $R_n = O(\sqrt{nd \log(nK)})$ where

$K = |\mathcal{A}|$. There is no contradiction because the action set in the above proof has $K = |\mathcal{A}| = 2^d$.

24.2 Sphere

Lower bounding the minimax regret when the action-set is the sphere presents an additional challenge relative to the hypercube. The product structure of the hypercube means that the learner can treat the dimensions independently, which is reflected in the lower bound. For the sphere this is not true because the magnitude of the action in one dimension constrains the learner in other dimensions. Nevertheless, almost the same technique with one modification allows us to prove a similar bound.

THEOREM 24.2 *Assume $d \leq 2n$ and let $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. Then there exists a $\theta \in \mathbb{R}^d$ with $\|\theta\|_2^2 = d^2/(4n)$ such that $R_n(\mathcal{A}, \theta) \geq d\sqrt{n}/16$.*

Proof Let $\Delta = \frac{1}{4}\sqrt{d/n}$ and $\theta \in \{\pm\Delta\}^d$ and $\tau_i = n \wedge \min\{t : \sum_{s=1}^t A_{si}^2 \geq n/d\}$. Then

$$\begin{aligned} R_n(\mathcal{A}, \theta) &= \Delta \mathbb{E}_\theta \left[\sum_{t=1}^n \sum_{i=1}^d \left(\frac{1}{\sqrt{d}} - A_{ti} \text{sign}(\theta_i) \right) \right] \\ &= \frac{\Delta\sqrt{d}}{2} \mathbb{E}_\theta \left[\sum_{t=1}^n \sum_{i=1}^d \left(\frac{1}{\sqrt{d}} - A_{ti} \text{sign}(\theta_i) \right)^2 \right] \\ &\geq \frac{\Delta\sqrt{d}}{2} \sum_{i=1}^d \mathbb{E}_\theta \left[\sum_{t=1}^{\tau_i} \left(\frac{1}{\sqrt{d}} - A_{ti} \text{sign}(\theta_i) \right)^2 \right]. \end{aligned}$$

Let $U_i(x) = \sum_{t=1}^{\tau_i} (1/\sqrt{d} - A_{ti}x)^2$ and $\theta' \in \{\pm\Delta\}^d$ be another parameter vector such that $\theta_j = \theta'_j$ for $j \neq i$ and $\theta'_i = -\theta_i$ and assume without loss of generality that $\theta_i > 0$. Let \mathbb{P} and \mathbb{P}' be the laws of $U_i(1)$ with respect to the bandit/learner interaction measure induced by θ and θ' respectively, then

$$\begin{aligned} \mathbb{E}_\theta[U_i(1)] &\geq \mathbb{E}_{\theta'}[U_i(1)] - \left(\frac{4n}{d} + 2 \right) \sqrt{\frac{1}{2} D(\mathbb{P}, \mathbb{P}')} \\ &\geq \mathbb{E}_{\theta'}[U_i(1)] - \frac{\Delta}{2} \left(\frac{4n}{d} + 2 \right) \sqrt{\mathbb{E} \left[\sum_{t=1}^{\tau_i} A_{ti}^2 \right]} \\ &\geq \mathbb{E}_{\theta'}[U_i(1)] - \frac{\Delta}{2} \left(\frac{4n}{d} + 2 \right) \sqrt{\frac{n}{d}} \\ &\geq \mathbb{E}_{\theta'}[U_i(1)] - \frac{4\Delta n}{d} \sqrt{\frac{n}{d}}, \end{aligned}$$

where in the first inequality we used Pinsker's inequality (Eq. (14.8)) and the

bound $U_i(1) \leq 4n/d + 2$, which follows from the definition of τ_i and the fact that $|A_{\tau_i i}| \leq 1$. In the second line we used the chain rule for the relative entropy up to a stopping time (Exercise 15.6). The second last inequality is true by the definition of τ_i and the last by the assumption that $d \leq 2n$.

$$\begin{aligned} \mathbb{E}_\theta[U_i(1)] + \mathbb{E}_{\theta'}[U_i(-1)] &\geq \mathbb{E}_{\theta'}[U_i(1) + U_i(-1)] - \frac{4n\Delta}{d}\sqrt{\frac{n}{d}} \\ &= 2\mathbb{E}_{\theta'}\left[\frac{\tau_i}{d} + \sum_{t=1}^{\tau_i} A_{ti}^2\right] - \frac{4n\Delta}{d}\sqrt{\frac{n}{d}} \geq \frac{2n}{d} - \frac{4n\Delta}{d}\sqrt{\frac{n}{d}} = \frac{n}{d}. \end{aligned}$$

The proof is completed using the randomization hammer:

$$\begin{aligned} \sum_{\theta \in \{\pm\Delta\}^d} R_n(\mathcal{A}, \theta) &\geq \frac{\Delta\sqrt{d}}{2} \sum_{i=1}^d \sum_{\theta \in \{\pm\Delta\}^d} \mathbb{E}_\theta[U_i(\text{sign}(\theta_i))] \\ &= \frac{\Delta\sqrt{d}}{2} \sum_{i=1}^d \sum_{\theta_{-i} \in \{\pm\Delta\}^{d-1}} \sum_{\theta_i \in \{\pm\Delta\}} \mathbb{E}_\theta[U_i(\text{sign}(\theta_i))] \\ &\geq \frac{\Delta\sqrt{d}}{2} \sum_{i=1}^d \sum_{\theta_{-i} \in \{\pm\Delta\}^{d-1}} \frac{n}{d} = 2^{d-2}n\Delta\sqrt{d}. \end{aligned}$$

Hence there exists a $\theta \in \{\pm\Delta\}^d$ such that $R_n(\mathcal{A}, \theta) \geq \frac{n\Delta\sqrt{d}}{4} = \frac{d\sqrt{n}}{16}$. \square

24.3 Sparse parameter vectors

In Chapter 23 we gave an algorithm with $R_n = \tilde{O}(\sqrt{dpn})$ where $p \geq \|\theta\|_0$ is a known bound on the sparsity of the unknown parameter. Except for logarithmic terms this bound cannot be improved. An extreme case is when $p = 1$, which essentially reduces to the finite-armed bandit problem where the minimax regret has order \sqrt{dn} (see Chapter 15). For this reason we cannot expect too much from sparsity and in particular the worst case bound will depend on polynomially on the ambient dimension d .

Constructing a lower bound for $p > 1$ is relatively straightforward. For simplicity we assume that $d = pk$ for some integer $k > 1$. A sparse linear bandit can mimic the learner playing p finite-armed bandits simultaneously, each with k arms. Rather than observing the reward for each bandit, however, the learner only observes the sum of the rewards and the noise is added at the end. This is sometimes called the **multitask bandit** problem.

THEOREM 24.3 *Assume $pd \leq n$ and there exists a natural number $k > 1$ such that $d = pk$. Let $\mathcal{A} = \{e_i : i \in [k]\}^p \subset \mathbb{R}^d$. Then for any policy there exists a $\theta \in \mathbb{R}^d$ with $\|\theta\|_0 = p$ and $\|\theta\|_\infty \leq \sqrt{d/(pn)}$ such that $R_n(\mathcal{A}, \theta) \geq \frac{1}{8}\sqrt{pdn}$.*

Proof Let $\Delta > 0$ and $\Theta = \{\Delta e_i : i \in [k]\} \subset \mathbb{R}^k$. Given $\theta \in \Theta^p$ and $i \in [p]$ let $\theta^{(i)} \in \mathbb{R}^k$ be defined by $\theta_k^{(i)} = \theta_{(i-1)p+k}$, which means that

$$\theta^\top = [\theta^{(1)\top}, \theta^{(2)\top}, \dots, \theta^{(p)\top}].$$

Next define matrix $V \in \mathbb{R}^{p \times d}$ be the matrix with $V_{ij} = 1 + (j-1) \bmod k$. For example, when $p=2$:

$$V = \begin{bmatrix} 1 & \cdots & k & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & k & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & k \end{bmatrix}.$$

Let $B_t = VA_t \in [k]^p$ represent the vector of ‘base’ actions chosen by the learner in each of the p bandits in round t . The optimal action in the i th bandit is

$$b_i^*(\theta) = \operatorname{argmax}_{b \in [k]} \theta_b^{(i)}.$$

The regret can be decomposed into

$$R_n(\theta) = \sum_{i=1}^p \underbrace{\Delta \mathbb{E}_\theta \left[\sum_{t=1}^n \mathbb{I}\{B_{ti} \neq b_i^*\} \right]}_{R_{ni}(\theta)}.$$

For $i \in [p]$ we abbreviate $\theta^{(-i)} = \theta^{(1)}, \dots, \theta^{(i-1)}, \theta^{(i+1)}, \dots, \theta^{(p)}$. Then

$$\begin{aligned} \frac{1}{|\Theta|^p} \sum_{\theta \in \Theta^p} R_n(\theta) &= \frac{1}{|\Theta|^p} \sum_{i=1}^p R_{ni}(\theta) \\ &= \sum_{i=1}^p \frac{1}{|\Theta|^{p-1}} \sum_{\theta^{(-i)} \in \Theta^{p-1}} \frac{1}{|\Theta|} \sum_{\theta^{(i)} \in \Theta} R_{ni}(\theta) \\ &\geq \frac{1}{8} \sum_{i=1}^p \frac{1}{|\Theta|^{p-1}} \sum_{\theta^{(-i)} \in \Theta^{p-1}} \sqrt{kn} \\ &= \frac{1}{8} p \sqrt{kn} = \frac{1}{8} \sqrt{dpn}. \end{aligned} \tag{24.3}$$

The only tricky step is the inequality, which follows by choosing $\Delta \approx \sqrt{k/n}$ and repeating the argument outlined in Exercise 15.1. We leave it to the reader to check the details (Exercise 24.1). \square

24.4 Unrealizable case

An important generalization of the linear model is the **unrealizable** case where the mean rewards are not assumed to follow a linear model exactly. Suppose that $\mathcal{A} \subset \mathbb{R}^d$ is a finite set with $|\mathcal{A}| = K$ and that $X_t = \eta_t + \mu(A_t)$ where $\mu : \mathcal{A} \rightarrow \mathbb{R}$ is an unknown function. Let $\theta \in \mathbb{R}^d$ be the parameter vector for which

$\sup_{a \in \mathcal{A}} |\langle \theta, a \rangle - \mu(a)|$ is as small as possible:

$$\theta = \operatorname{argmin}_{\theta \in \mathbb{R}^d} \sup_{a \in \mathcal{A}} |\langle \theta, a \rangle - \mu(a)| .$$

Then let $\varepsilon = \sup_{a \in \mathcal{A}} |\langle \theta, a \rangle - \mu(a)|$ be the maximum error. It would be very pleasant to have an algorithm such that

$$R_n(\mathcal{A}, \mu) = n \max_{a \in \mathcal{A}} \mu(a) - \mathbb{E} \left[\sum_{t=1}^n \mu(A_t) \right] = \tilde{O}(\min\{d\sqrt{n} + \varepsilon n\}, \sqrt{Kn}) . \quad (24.4)$$

Unfortunately it turns out that results of this kind are not achievable. To show this we will prove a generic bound for the classical finite-armed bandit problem and afterwards show how this implies the impossibility of an adaptive bound like the above.

THEOREM 24.4 *Let $\mathcal{A} = [K]$ and for $\mu \in [0, 1]^K$ the reward is $X_t = \mu_{A_t} + \eta_t$ and the regret is*

$$R_n(\mu) = n \max_{i \in \mathcal{A}} \mu_i - \mathbb{E}_\mu \left[\sum_{t=1}^n \mu_{A_t} \right] .$$

Define $\Theta, \Theta' \subset \mathbb{R}^K$ by

$$\Theta = \{\mu \in [0, 1]^K : \mu_i = 0 \text{ for } i > 1\} \quad \Theta' = \{\mu \in [0, 1]^K\} .$$

If $V \in \mathbb{R}$ is such that $2(K-1) \leq V \leq \sqrt{n(K-1)\exp(-2)/8}$ and $\sup_{\mu \in \Theta} R_n(\mu) \leq V$, then

$$\sup_{\mu' \in \Theta'} R_n(\mu') \geq \frac{n(K-1)}{8V} \exp(-2) .$$

Proof Recall that $T_i(n) = \sum_{t=1}^n \mathbb{I}\{A_t = i\}$ is the number of times arm i is played after all n rounds. Let $\mu \in \Theta$ be given by $\mu_1 = \Delta = (K-1)/V \leq 1/2$. The regret is then decomposed as:

$$R_n(\mu) = \Delta \sum_{i=2}^K \mathbb{E}_\mu[T_i(n)] \leq V .$$

Rearranging shows that $\sum_{i=2}^K \mathbb{E}_\mu[T_i(n)] \leq \frac{V}{\Delta}$ and so by the pigeonhole principle there exists an $i > 1$ such that

$$\mathbb{E}_\mu[T_i(n)] \leq \frac{V}{(K-1)\Delta} = \frac{1}{\Delta^2} .$$

Then define $\mu' \in \Theta'$ by

$$\mu'_j = \begin{cases} \Delta & \text{if } j = 1 \\ 2\Delta & \text{if } j = i \\ 0 & \text{otherwise.} \end{cases}$$

Then by Theorem 14.2 and Lemma 15.1, for any event A we have

$$\mathbb{P}_\mu(A) + \mathbb{P}_{\mu'}(A^c) \geq \frac{1}{2} \exp(D(\mathbb{P}_\mu, \mathbb{P}_{\mu'})) = \frac{1}{2} \exp(-2\Delta^2 \mathbb{E}[T_i(n)]) \geq \frac{1}{2} \exp(-2).$$

By choosing $A = \{T_1(n) \leq n/2\}$ we have

$$R_n(\mu) + R_n(\mu') \geq \frac{n\Delta}{4} \exp(-2) = \frac{n(K-1)}{4V} \exp(-2).$$

Therefore by the assumption that $R_n(\mu) \leq V \leq \sqrt{n(K-1)\exp(-2)/8}$ we have

$$R_n(\mu') \geq \frac{n(K-1)}{8V} \exp(-2).$$

By the definition of V we conclude that $R_n(\mu)R_n(\mu') \geq \frac{n(K-1)}{8} \exp(-2)$ as required. \square

As promised we now relate this to the unrealizable linear bandits. Suppose that $d = 1$ (an absurd case) and that there are K arms $\mathcal{A} = \{a_1, a_2, \dots, a_K\} \subset \mathbb{R}^1$ where $a_1 = (1)$ and $a_i = (0)$ for $i > 1$. Clearly if $\theta > 0$ and $\mu(a_i) = \langle a_i, \theta \rangle$, then the problem can be modelled as a finite-armed bandit with means $\mu \in \Theta \subset [0, 1]^K$. In the general case we just have a finite-armed bandit with $\mu \in \Theta'$. If in the first case we have $R_n(\mathcal{A}, \mu) = O(\sqrt{n})$, then the theorem shows for large enough n that

$$\sup_{\mu' \in \Theta'} R_n(\mathcal{A}, \mu) = O(K\sqrt{n}).$$

It follows that Eq. (24.4) is a pipe dream. To our knowledge it is still an open question of what is possible on this front. Our conjecture is that there is a policy for which

$$R_n(\mathcal{A}, \theta) = \tilde{O}\left(\min\left\{d\sqrt{n} + \varepsilon n, \frac{K}{d}\sqrt{n}\right\}\right).$$

In fact, it is not hard to design an algorithm that tries to achieve this bound by assuming the problem is realizable, but using some additional time to explore the remaining arms up to some accuracy to confirm the hypothesis.

24.5 Notes

- 1 The worst-case bound demonstrates the near-optimality of the OFUL algorithm for a specific action set. It is an open question to characterize the optimal regret for a wide range of action sets. We will return to these issues soon when we discuss adversarial linear bandits.

24.6 Bibliographic remarks

Worst-case lower bounds for stochastic bandits have appeared in a variety of places, all with roughly the same bound, but for different action sets. Our very simple

proof for the hypercube is new, but takes inspiration from the paper by [Shamir \[2015\]](#). The first lower bound for the sphere was given by [Rusmevichientong and Tsitsiklis \[2010\]](#) with smaller constants and a complicated proof. As far as we know the first lower bound of $\Omega(d\sqrt{n})$ was given by [Dani et al. \[2008\]](#) for an action-set equal to the product of 2-dimensional disks. The results for the unrealizable case are inspired by the work of one of the authors on the Pareto-regret frontier for bandits, which characterizes what trade-offs are available when it is desirable to have a regret that is unusually small relative to some specific arms [\[Lattimore, 2015a\]](#).

24.7 Exercises

24.1 Completing the missing steps to prove the inequality in Eq. (24.3).

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

25 Asymptotic Lower Bounds for Stochastic Linear Bandits

The lower bounds in the previous chapter were derived by analyzing the worst case for specific action sets and/or constraints on the unknown parameter. In this chapter we focus on the asymptotics of the problem and aim to understand the influence of the action set on the regret. We assume that $\mathcal{A} \subset \mathbb{R}^d$ is finite with $|\mathcal{A}| = K$ and that the reward is $X_t = \langle A_t, \theta \rangle + \eta_t$ where $\theta \in \mathbb{R}^d$ and η_t is a sequence of independent standard Gaussian noise. Of course the regret of a policy in this setting is

$$R_n(\mathcal{A}, \theta) = \mathbb{E}_\theta \left[\sum_{t=1}^n \Delta_{A_t} \right], \quad \Delta_a = \max_{a' \in \mathcal{A}} \langle a' - a, \theta \rangle,$$

where the dependence on the policy is omitted for readability and $\mathbb{E}_\theta[\cdot]$ is the expectation with respect to the measure on outcomes induced by the interaction of the policy and the linear bandit determined by θ . Like the asymptotic lower bounds in the classical finite-armed case (Chapter 16), the results of this chapter are proven only for consistent policies. Recall that a policy is consistent in some class of bandits \mathcal{E} if the regret is subpolynomial for any bandit in that class. Here this means that

$$R_n(\mathcal{A}, \theta) = o(n^p) \quad \text{for all } p > 0 \text{ and } \theta \in \mathbb{R}^d. \quad (25.1)$$

The main objective of the chapter is to prove the following theorem on the behaviour of any consistent policy and discuss the implications.

THEOREM 25.1 *Assume that $\mathcal{A} \subset \mathbb{R}^d$ is finite and spans \mathbb{R}^d and suppose a policy is consistent (satisfies Eq. 25.1). Let $\theta \in \mathbb{R}^d$ be any parameter such that there is a unique optimal action and let $\bar{G}_n = \mathbb{E}_\theta [\sum_{t=1}^n A_t A_t^\top]$ be the expected Gram matrix. Then $\liminf_{n \rightarrow \infty} \lambda_{\min}(\bar{G}_n) / \log(n) > 0$. Furthermore, for any $a \in \mathcal{A}$ it holds that:*

$$\limsup_{n \rightarrow \infty} \log(n) \|a\|_{\bar{G}_n^{-1}}^2 \leq \frac{\Delta_a^2}{2}.$$

The reader should recognize $\|a\|_{\bar{G}_n^{-1}}^2$ as the key term in the width of the confidence interval for the least squares estimator (Chapter 20). This is quite intuitive. The theorem is saying that any consistent algorithm must prove statistically that all suboptimal arms are indeed suboptimal by making the size of the confidence interval smaller than the suboptimality gap. Before the

proof of this result we give a corollary that characterizes the asymptotic regret that must be endured by any consistent policy.

COROLLARY 25.1 *Let $\mathcal{A} \subset \mathbb{R}^d$ be a finite set that spans \mathbb{R}^d and $\theta \in \mathbb{R}^d$ be such that there is a unique optimal action. Then for any consistent policy*

$$\liminf_{n \rightarrow \infty} \frac{R_n(\mathcal{A}, \theta)}{\log(n)} \geq c(\mathcal{A}, \theta),$$

where $c(\mathcal{A}, \theta)$ is defined as

$$c(\mathcal{A}, \theta) = \inf_{\alpha \in [0, \infty)^{\mathcal{A}}} \sum_{a \in \mathcal{A}} \alpha(a) \Delta_a$$

subject to $\|a\|_{H_\alpha^{-1}}^2 \leq \frac{\Delta_a^2}{2}$ for all $a \in \mathcal{A}$ with $\Delta_a > 0$,

where $H = \sum_{a \in \mathcal{A}} \alpha(a) aa^\top$.

The lower bound is complemented by a matching upper bound that we will not prove.

THEOREM 25.2 *Let $\mathcal{A} \subset \mathbb{R}^d$ be a finite set that spans \mathbb{R}^d . Then there exists a policy such that*

$$\limsup_{n \rightarrow \infty} \frac{R_n(\mathcal{A}, \theta)}{\log(n)} \leq c(\mathcal{A}, \theta),$$

where \mathcal{A} is defined as in Corollary 25.1.

Proof of Theorem 25.1 The proof of the first part is simply omitted (see the reference below for details). It follows along similar lines to what follows, essentially that if G_n is not sufficiently large in every direction, then some alternative parameter is not sufficiently identifiable. Let $a^* = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \theta \rangle$ be the optimal action, which we assumed to be unique. Let $\theta' \in \mathbb{R}^d$ be an alternative parameter to be chosen subsequently and let \mathbb{P} and \mathbb{P}' be the measures on the sequence of outcomes $A_1, Y_1, \dots, A_n, Y_n$ induced by the interaction between the policy and the bandit determined by θ and θ' respectively. Let $\mathbb{E}[\cdot]$ and $\mathbb{E}'[\cdot]$ be the expectation operators of \mathbb{P} and \mathbb{P}' respectively. By Theorem 14.2 and Lemma 15.1 for any event E we have

$$\begin{aligned} \mathbb{P}(E) + \mathbb{P}'(E^c) &\geq \frac{1}{2} \exp(-D(\mathbb{P}, \mathbb{P}')) \\ &= \frac{1}{2} \exp\left(-\frac{1}{2} \mathbb{E}\left[\sum_{t=1}^n \langle A_t, \theta - \theta' \rangle^2\right]\right) = \frac{1}{2} \exp\left(-\frac{1}{2} \|\theta - \theta'\|_{G_n}^2\right). \end{aligned} \tag{25.2}$$

A simple re-arrangement shows that

$$\frac{1}{2} \|\theta - \theta'\|_{G_n}^2 \geq \log\left(\frac{1}{2\mathbb{P}(E) + 2\mathbb{P}'(E^c)}\right).$$

Now we follow the usual plan of choosing θ' to be close to θ , but so that the

optimal action in the bandit determined by θ' is not a^* . Let $\Delta_{\min} = \min\{\Delta_a : a \in \mathcal{A}, \Delta_a > 0\}$ and $\varepsilon \in (0, \Delta_{\min})$ and H be a positive definite matrix to be chosen later such that $\|a - a^*\|_H^2 > 0$. Then define

$$\theta' = \theta + \frac{\Delta_a + \varepsilon}{\|a - a^*\|_H^2} H(a - a^*),$$

which is chosen so that

$$\langle a - a^*, \theta' \rangle = \langle a - a^*, \theta \rangle + \Delta_a + \varepsilon = \varepsilon.$$

This means that a^* is ε -suboptimal action for bandit θ' . We abbreviate $R_n = R_n(\mathcal{A}, \theta)$ and $R'_n = R_n(\mathcal{A}, \theta')$. Then

$$R_n = \mathbb{E} \left[\sum_{a \in \mathcal{A}} T_a(n) \Delta_a \right] \geq \frac{n \Delta_{\min}}{2} \mathbb{P}(T_{a^*}(n) < n/2) \geq \frac{n \varepsilon}{2} \mathbb{P}(T_{a^*}(n) < n/2),$$

where $T_a(n) = \sum_{t=1}^n \mathbb{I}\{A_t = a\}$. Similarly, a^* is at least ε -suboptimal in bandit θ' so that

$$R'_n \geq \frac{n \varepsilon}{2} \mathbb{P}'(T_{a^*}(n) \geq n/2).$$

Therefore

$$\mathbb{P}(T_{a^*}(n) < n/2) + \mathbb{P}'(T_{a^*}(n) \geq n/2) \leq \frac{2}{n \varepsilon} (R_n + R'_n). \quad (25.3)$$

Note that this holds for practically any choice of H as long as $\|a - a^*\|_H > 0$. The logical next step is to select H (which determines θ') to make (25.2) as large as possible. The main difficulty is that this depends on n , so instead we aim to choose an H so the quantity is large enough infinitely often. We start by just re-arranging things:

$$\frac{1}{2} \|\theta - \theta'\|_{\bar{G}_n}^2 = \frac{(\Delta_a + \varepsilon)^2}{2} \cdot \frac{\|a - a^*\|_{H\bar{G}_n H}^2}{\|a - a^*\|_H^4} = \frac{(\Delta_a + \varepsilon)^2}{2\|a - a^*\|_{\bar{G}_n^{-1}}^2} \rho_n(H),$$

where we introduced

$$\rho_n(H) = \frac{\|a - a^*\|_{\bar{G}_n^{-1}}^2 \|a - a^*\|_{H\bar{G}_n H}^2}{\|a - a^*\|_H^4}.$$

Therefore by choosing E to be the event that $T_{a^*}(n) < n/2$ and using (25.3) and (25.2) we have

$$\frac{(\Delta_a + \varepsilon)^2}{2\|a - a^*\|_{\bar{G}_n^{-1}}^2} \rho_n(H) \geq \log \left(\frac{n \varepsilon}{4R_n + 4R'_n} \right),$$

which after re-arrangement leads to

$$\frac{(\Delta_a + \varepsilon)^2}{2 \log(n) \|a - a^*\|_{\bar{G}_n^{-1}}^2} \rho_n(H) \geq 1 - \frac{\log((4R_n + 4R'_n)/\varepsilon)}{\log(n)}.$$

The definition of consistency means that R_n and R'_n are both sub-polynomial,

which implies that the second term in the previous expression tends to zero for large n and so by sending ε to zero we see that

$$\liminf_{n \rightarrow \infty} \frac{\rho_n(H)}{\log(n) \|a - a^*\|_{\bar{G}_n^{-1}}^2} \geq \frac{2}{\Delta_a^2}. \quad (25.4)$$

We complete the result using proof by contradiction. Suppose that

$$\limsup_{n \rightarrow \infty} \log(n) \|a - a^*\|_{\bar{G}_n^{-1}}^2 > \frac{\Delta_a^2}{2}. \quad (25.5)$$

Then there exists an $\varepsilon > 0$ and infinite set $S \subseteq \mathbb{N}$ such that

$$\log(n) \|a - a^*\|_{\bar{G}_n^{-1}}^2 \geq \frac{(\Delta_a + \varepsilon)^2}{2} \quad \text{for all } n \in S.$$

Therefore by (25.4), $\liminf_{n \in S} \rho_n(H) > 1$. We now choose H to be a cluster point of the sequence $(\bar{G}_n^{-1} / \|\bar{G}_n^{-1}\|)_{n \in S}$ where $\|\bar{G}_n^{-1}\|$ is the spectral norm of the matrix \bar{G}_n^{-1} . Such a point must exist, since matrices in this sequence have unit spectral norm by definition, and the set of matrices with bounded spectral norm is compact. We let $S' \subseteq S$ be a subset so that $\bar{G}_n^{-1} / \|\bar{G}_n^{-1}\|$ converges to H on $n \in S'$. We now check that $\|a - a^*\|_H > 0$.

$$\|a - a^*\|_H^2 = \lim_{n \in S} \frac{\|a - a^*\|_{\bar{G}_n^{-1}}^2}{\|\bar{G}_n^{-1}\|} > 0,$$

where the last inequality follows from the assumption in (25.5) and the first part of the theorem. Therefore

$$1 < \liminf_{n \in S} \rho_n(H) \leq \liminf_{n \in S'} \frac{\|a - a^*\|_{\bar{G}_n^{-1}}^2 \|a - a^*\|_{H\bar{G}_n^{-1}H}^2}{\|a - a^*\|_H^4} = 1,$$

which is a contradiction, and so we conclude that (25.5) does not hold and so

$$\limsup_{n \rightarrow \infty} \log(n) \|a - a^*\|_{\bar{G}_n^{-1}}^2 \leq \frac{\Delta_a^2}{2}. \quad \square$$

We leave the proof of the corollary as an exercise for the reader. Essentially though, any consistent algorithm must choose its actions so that in expectation

$$\|a - a^*\|_{\bar{G}_n^{-1}}^2 \leq (1 + o(1)) \frac{\Delta_a^2}{2 \log(n)}.$$

Now since a^* will be chosen linearly often it is easily shown for suboptimal a that $\lim_{n \rightarrow \infty} \|a - a^*\|_{\bar{G}_n^{-1}} / \|a\|_{\bar{G}_n^{-1}} \rightarrow 1$. This leads to the required constraint on the actions of the algorithm, and the optimization problem in the corollary is derived by minimizing the regret subject to this constraint.

25.1 Clouds looming for optimism

The theorem and its corollary have disturbing implications for policies based on the principle of optimism in the face of uncertainty, which is that they can never be asymptotically optimal. The reason is that these policies do not choose actions for which they have collected enough statistics to prove they are suboptimal, but in the linear setting it can still be worthwhile playing these actions in case they are very informative about other actions for which the statistics are not yet so clear. As we shall see, a problematic example appears in the simplest case where there is information sharing between the arms. Namely, when the dimension is $d = 2$ and there are $K = 3$ arms.

Let $\mathcal{A} = \{a_1, a_2, a_3\}$ where $a_1 = e_1$ and $a_2 = e_2$ and $a_3 = (1 - \varepsilon, \gamma\varepsilon)$ where $\gamma \geq 1$ and $\varepsilon > 0$ is small. Let $\theta = (1, 0)$ so that the optimal action is $a^* = a_1$ and $\Delta_{a_2} = 1$ and $\Delta_{a_3} = \varepsilon$. Clearly if ε is very small, then a_1 and a_3 point in nearly the same direction and so choosing only these arms does not provide sufficient information to quickly learn which of a_1 or a_3 is optimal. On the other hand, a_2 and $a_1 - a_3$ point in very different directions and so choosing a_2 allows a learning agent to quickly identify that a_1 is in fact optimal. We now show how the theorem and corollary demonstrate this. First we calculate what is the optimal solution to the optimization problem in Corollary 25.1. Recall we are trying to minimize

$$\sum_{a \in \mathcal{A}} \alpha(a) \Delta_a \quad \text{subject to } \|a\|_{H(\alpha)^{-1}}^2 \leq \frac{\Delta_a^2}{2} \text{ for all } a \in \mathcal{A},$$

where $H = \sum_{a \in \mathcal{A}} \alpha(a) aa^\top$. Clearly we should choose $\alpha(a_1)$ arbitrarily large, then a computation shows that

$$\lim_{\alpha(a_1) \rightarrow \infty} H(\alpha)^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\alpha(a_3)\varepsilon^2\gamma^2 + \alpha(a_2)} \end{bmatrix}.$$

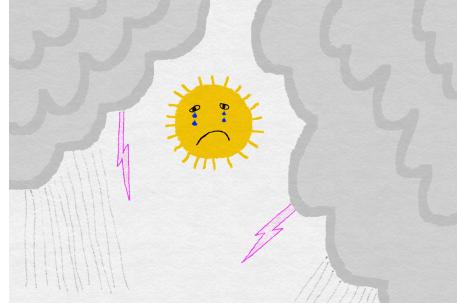
The constraints mean that

$$\begin{aligned} \frac{1}{\alpha(a_3)\varepsilon^2\gamma^2 + \alpha(a_2)} &= \lim_{\alpha(a_1) \rightarrow \infty} \|a_2\|_{H(\alpha)^{-1}}^2 \leq \frac{1}{2} \\ \frac{\gamma^2\varepsilon^2}{\alpha(a_3)\varepsilon^2\gamma^2 + \alpha(a_2)} &= \lim_{\alpha(a_1) \rightarrow \infty} \|a_3\|_{H(\alpha)^{-1}}^2 \leq \frac{\varepsilon^2}{2}. \end{aligned}$$

Provided that $\gamma \geq 1$ this reduces simply to the constraint that

$$\alpha(a_3)\varepsilon^2 + \alpha(a_2) \geq 2\gamma^2.$$

Since we are minimizing $\alpha(a_2) + \varepsilon\alpha(a_3)$ we can easily see that $\alpha(a_2) = 2\gamma^2$ and $\alpha(a_3) = 0$ provided that $2\gamma^2 \leq 2/\varepsilon$. Therefore if ε is chosen sufficiently small relative to γ , then the optimal rate of the regret is $c(\mathcal{A}, \theta) = 2\gamma^2$ and so by



Theorem 25.2 there exists a policy such that

$$\limsup_{n \rightarrow \infty} \frac{R_n(\mathcal{A}, \theta)}{\log(n)} = 2\gamma^2.$$

Now we argue that for γ sufficiently large and ε arbitrarily small that the regret for any consistent optimistic algorithm is at least

$$\limsup_{n \rightarrow \infty} \frac{R_n(\mathcal{A}, \theta)}{\log(n)} = \Omega(1/\varepsilon),$$

which can be arbitrarily worse than the optimal rate! So why is this so? Recall that optimistic algorithms choose

$$A_t = \operatorname{argmax}_{a \in \mathcal{A}} \max_{\tilde{\theta} \in \mathcal{C}_t} \langle a, \tilde{\theta} \rangle,$$

where $\mathcal{C}_t \subset \mathbb{R}^d$ is a confidence set that we assume contains the true θ with high probability. So far this does not greatly restrict the class of algorithms that we might call optimistic. We now assume that there exists a constant $c > 0$ such that

$$\mathcal{C}_t \subseteq \left\{ \tilde{\theta} : \|\hat{\theta}_t - \tilde{\theta}\|_{G_t} \leq c\sqrt{\log(n)} \right\}.$$

So now we ask how often can we expect the optimistic algorithm to choose action $a_2 = e_2$ in the example described above? Since we have assumed $\theta \in \mathcal{C}_t$ with high probability we have that

$$\max_{\tilde{\theta} \in \mathcal{C}_t} \langle a_1, \tilde{\theta} \rangle \geq 1.$$

On the other hand, if $T_{a_2}(t-1) > 4c^2 \log(n)$, then

$$\max_{\tilde{\theta} \in \mathcal{C}_t} \langle a_2, \tilde{\theta} \rangle = \max_{\tilde{\theta} \in \mathcal{C}_t} \langle a_2, \tilde{\theta} - \theta \rangle \leq 2c\sqrt{\|a_2\|_{G_t^{-1}} \log(n)} \leq 2c\sqrt{\frac{\log(n)}{T_{a_2}(t-1)}} < 1,$$

which means that a_2 will not be chosen more than $1 + 4c^2 \log(n)$ times. So if $\gamma = \Omega(c^2)$, then the optimistic algorithm will not choose a_2 sufficiently often and a simple computation shows it must choose a_3 at least $\Omega(\log(n)/\varepsilon^2)$ times and suffers regret of $\Omega(\log(n)/\varepsilon)$. The key take-away from this is that optimistic algorithms do not choose actions that are statistically suboptimal, but for linear bandits it can be optimal to choose these actions more often to gain information about *other actions*.

25.2 Notes

- 1 The algorithm that realizes Theorem 25.2 is a complicated three-phase affair that we cannot recommend in practice. A practical asymptotically optimal algorithm for linear bandits is a fascinating open problem.

- 2 In Chapter 35 we will introduce the randomized Bayesian algorithm called Thompson sampling algorithm for finite-armed and linear bandits. While Thompson sampling comes with several benefits over UCB, it does not overcome the issues described here.
- 3 The main difficulty in designing asymptotically optimal algorithms is how to balance the tradeoff between information and regret. One algorithm that tries to do this in an explicit way is “Information-Directed Sampling” by Russo and Roy [2014a], which we also discuss in Chapter 35. It is not known if the algorithm proposed there is optimal when adapted to linear bandits.

25.3 Bibliographic remarks

The theorems of this chapter are by the authors: Lattimore and Szepesvári [2017]. The example in Section 25.1 first appeared in a paper by Soare et al. [2014], which deals with the problem of best arm identification for linear bandits (for an introduction to best arm identification see Chapter 33).

25.4 Exercises

25.1 Prove Corollary 25.1.

25.2 Prove the first part of Theorem 25.1.

25.3 Give an example of an action set $\mathcal{A} \subset \mathbb{R}^d$ and $\theta \in \mathbb{R}^d$ and vector $a \in \mathbb{R}^d$ where the asymptotic regret for the same θ and action-set $\mathcal{A} \cup \{a\}$ is:

- (a) Makes the asymptotic regret larger.
- (b) Makes the asymptotic regret smaller.

Part VI

Adversarial Linear Bandits

In the next few chapters we will consider adversarial linear bandits, which superficially can be thought of as the adversarial version of the stochastic linear bandit. Indeed, the techniques in this chapter combine the ideas of optimal design presented in Chapter 22 with the exponential weighting algorithm of Chapter 11. The intuitions gained by studying stochastic bandits should not be taken too seriously here, however. There are subtle differences between the model of adversarial bandits introduced here and the stochastic linear bandits examined in previous chapters. These differences will be discussed at length in Chapter 29. The adversarial version of the linear bandits turns out to be remarkably rich, both because of the complex information structure and because of the challenging computational issues.

The part is split into four chapters, the first of which is an introduction to the necessary tools from convex analysis and optimization. In the first chapter on bandits we show how to combine the core ideas of the Exp3 policy of Chapter 11 with the optimal experimental design for least squares estimators in Chapter 21. When the number of actions is large (or infinite), the approach based on Exp3 is hard to make efficient. These shortcomings are addressed in the next chapter where we introduce the mirror descent and follow-the-regularized leader algorithms for bandits and show how they can be used to design efficient algorithms. We conclude the part with a discussion on the relationship between adversarial and stochastic linear bandits, which is more subtle than the situation with finite-armed bandits.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

26 Foundations of Convex Analysis (†)

Our coverage of convexity is necessarily extremely brief. We introduce only what is necessary and refer the reader to standard texts for the proofs.

26.1 Convex sets and functions

A set $A \subseteq \mathbb{R}^d$ is convex if for any $x, y \in A$ it holds that $\alpha x + (1 - \alpha)y \in A$ for all $\alpha \in (0, 1)$. The convex hull of a collection of points $x_1, x_2, \dots, x_n \in \mathbb{R}^d$ is the smallest convex set containing the points, which also happens to satisfy

$$\text{co}(x_1, x_2, \dots, x_m) = \left\{ x \in \mathbb{R}^d : \text{exists } p \in \mathcal{P}_{d-1} \text{ such that } x = \sum_{i=1}^n p_i x_i \right\}.$$

The convex hull is also defined for an arbitrary set $A \subset \mathbb{R}^d$: $\text{co}(A)$, the convex hull of A is defined to be the smallest convex set containing A (see (c) in Figure 26.1). Let $A \subset \mathbb{R}^d$. Then the **polar** of A is denoted by A° and defined

$$A^\circ = \left\{ u \in \mathbb{R}^d : \sup_{x \in A} |\langle u, x \rangle| \leq 1 \right\}.$$

Of course if $\sup_{x \in A} |\langle x, u \rangle| \leq 1$ and $\sup_A |\langle x, v \rangle| \leq 1$, then $\sup_A |\langle x, \alpha u + (1 - \alpha)v \rangle| \leq 1$, which ensures that the polar is convex (even for non-convex A). For the rest of the section we let $A \subseteq \mathbb{R}^d$ be convex. Let $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ be the extended real number system and define operations involving infinities in the usual way (see notes). An extended real-valued function $f : \mathbb{R}^d \rightarrow \bar{\mathbb{R}}$ is **convex** if its **epigraph**

$$E_f = \{(x, y) : x \in \mathbb{R}^d, y \geq f(x)\} \subset \mathbb{R}^{d+1}$$

is a convex set. The **domain** of a convex function on \mathbb{R}^d is $\text{dom}(f) = \{x \in \mathbb{R}^d : f(x) < \infty\}$. A convex function is **proper** if its domain is nonempty and its range does not include $-\infty$.



For the rest of the chapter we will write “let f be a convex” to mean that $f : \mathbb{R}^d \rightarrow \bar{\mathbb{R}}$ is a proper convex function. Permitting convex functions to take values of $-\infty$ is a convenient standard because certain operations on proper

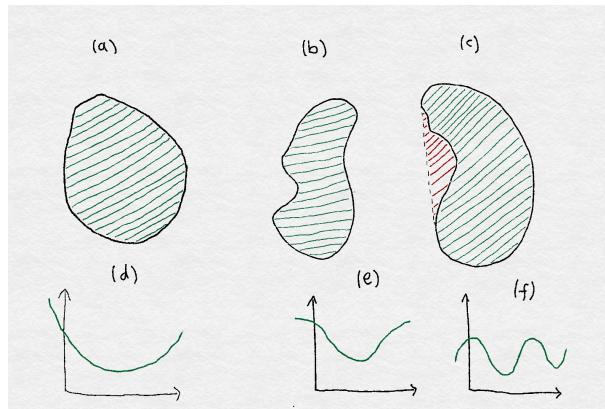


Figure 26.1 (a) is a convex set. (b) is a nonconvex set. (c) is the convex hull of a nonconvex set. (d) is a convex function. (e) is nonconvex, but all local minimums are global. (f) is not convex.

convex functions result in improper ones (infimal convolution, for example). These technicalities will never bother us in this book, however.

A consequence of the definition is that for convex f we have

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) \quad \text{for all } \alpha \in (0, 1) \text{ and } x, y \in \text{dom}(f). \quad (26.1)$$



Some authors use Eq. (26.1) as the definition of a convex function along with a specification that the domain is convex. If $A \subseteq \mathbb{R}^d$ is convex, then $f : A \rightarrow \mathbb{R}$ is convex if it satisfies Eq. (26.1).

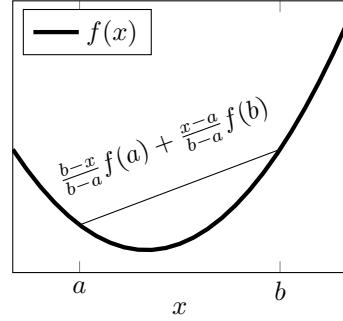
A function is **strictly convex** if the inequality in Eq. (26.1) is always strict. The **Fenchel dual** of a function f is $f^*(u) = \sup_{x \in \text{dom}(f)} \langle x, u \rangle - f(x)$, which is convex because the maximum of convex functions is convex. The Fenchel dual is also called the convex conjugate. If $f : A \rightarrow \mathbb{R}$ is twice differentiable, then convexity of f is equivalent to its Hessian having nonnegative eigenvalues for all $x \in A$. Strict convexity is equivalent to having strictly positive eigenvalues. The field of optimization is obsessed with convex functions because all local minimums are global (see figure). This means that minimizing a convex function is usually possible (efficiently) using some variation of gradient descent. A function $f : A \rightarrow \mathbb{R}$ is **concave** if $-f$ is convex.

26.2 Jensen's inequality

One of the most important results for convex functions is Jensen's inequality.

THEOREM 26.1 (Jensen's inequality) *Let (A, \mathcal{F}, P) be a probability space on convex set $A \subset \mathbb{R}^d$ and $f : A \rightarrow \mathbb{R}$ is \mathcal{F} -measurable and convex and $X(\omega) = \omega$ is the identity random element, then $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.*

Perhaps the archetypical application is that for any convex $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and $x, y \in \mathbb{R}^d$ we have $\alpha f(x) + (1 - \alpha)f(y) \geq f(\alpha x + (1 - \alpha)y)$ for all $\alpha \in [0, 1]$. Jensen's inequality is so central to convexity that it can actually be used as the definition (a function is convex if and only if it satisfies Jensen's inequality). The proof of Jensen's using the standard definition is not hard, but we include only a picture to convince the reader. The direction of Jensen's inequality is reversed if 'convex' is replaced by 'concave'.

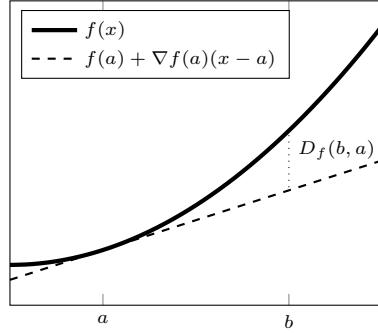


26.3 Bregman divergence

Let $f : A \rightarrow \mathbb{R}$ be convex and differentiable and let $x, y \in A$. Then the **Bregman divergence** induced by f is defined by

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle,$$

where $\nabla f(y) \in \mathbb{R}^d$ is the gradient of f at y . To get a sense of the divergence function D_f , note that $D_f(x, y)$ is the difference between $f(x)$ and its first order Taylor expansion about the point y . Since f is convex, the linear approximation of f is a lower bound on f and so $D_f(x, y)$ is nonnegative over its domain with $D_f(x, x) = 0$.



THEOREM 26.2 *The following hold:*

- (a) $D_f(x, y) \geq 0$ for all $x, y \in A$.
- (b) $D_f(x, x) = 0$ for all $x \in A$.
- (c) $D_f(x, y)$ is convex as a function of x .

The square root of the Bregman divergence shares many properties with a metric and for some choices of f it actually is a metric. In general, however, it is not symmetric and does not satisfy the triangle inequality.

EXAMPLE 26.1 Let $A = \mathbb{R}^d$ and $f(x) = \frac{1}{2}\|x\|_2^2$. Then $\nabla f(x) = x$ and

$$D_f(x, y) = \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x - y \rangle = \frac{1}{2}\|x - y\|_2^2.$$

EXAMPLE 26.2 Let $A = [0, \infty)^d$ and $f(x) = \sum_{i=1}^d (x_i \log(x_i) - x_i)$. Then $\nabla f(x) = \log(x)$ and

$$\begin{aligned} D_f(x, y) &= \sum_{i=1}^d (x_i \log(x_i) - x_i) - \sum_{i=1}^d (y_i \log y_i - y_i) - \sum_{i=1}^d \log(y_i)(x_i - y_i) \\ &= \sum_{i=1}^d x_i \log\left(\frac{x_i}{y_i}\right) + \sum_{i=1}^d (y_i - x_i). \end{aligned}$$

Notice that if $x, y \in \mathcal{P}_{d-1}$ are in the unit simplex, then $D_f(x, y)$ is the relative entropy between probability vectors x and y . The function f is called the **unnormalized negentropy**, which will feature heavily in many of the chapters that follow.

26.4 Legendre functions

In this section we use various topological notions such as the interior, closed set and boundary. The definitions of these terms are given in the notes. Let f be a convex function and $A = \text{dom}(f)$ and $C = \text{int}(A)$. Then f is **essentially smooth** if:

- (a) C is nonempty.
- (b) f is differentiable on C .
- (c) $\lim_{n \rightarrow \infty} \|\nabla f(x_n)\|_2 = \infty$ for any sequence $(x_n)_n$ with $x_n \in C$ for all n and $\lim_n x_n = x$ and some $x \in \partial C$.

It is **essentially strictly convex** if f is strictly convex on every convex subset of $\text{dom}(\nabla f)$. A **Legendre function** is a convex function f that is both essentially smooth and essentially strictly convex. The intuition is that the set $\{(x, f(x)) : x \in \text{dom}(f)\}$ is a ‘dish’ with ever-steepening edges towards the boundary of the domain.

THEOREM 26.3 *Let $f : \text{dom}(f) \rightarrow \mathbb{R}$ be a Legendre function. Then:*

- (a) ∇f is a bijection between $\text{int}(\text{dom}(f))$ and $\text{int}(\text{dom}(f^*))$ with the inverse $(\nabla f)^{-1} = \nabla f^*$.
- (b) $D_f(x, y) = D_{f^*}(\nabla f(y), \nabla f(x))$ for all $x, y \in \text{int}(\text{dom}(f))$.

The next corollary formalizes the ‘dish’ intuition by showing the directional derivative along any straight path from a point in the interior to the boundary blows up.

COROLLARY 26.1 *Let f be Legendre and $x \in \text{int}(\text{dom}(f))$ and $y \in \partial \text{int}(\text{dom}(f))$, then $\lim_{\alpha \rightarrow 1} \langle \nabla f((1-\alpha)x + \alpha y), y - x \rangle = \infty$.*

EXAMPLE 26.3 Let f be the Legendre function given by $f(x) = \frac{1}{2}\|x\|_2^2$, which has domain $\text{dom}(f) = \mathbb{R}^d$. Then $f^*(x) = f(x)$ and ∇f and ∇f^* are the identity functions.

EXAMPLE 26.4 Let $f(x) = -2 \sum_{i=1}^d \sqrt{x_i}$ when $x_i \geq 0$ for all i and ∞ otherwise, which has $\text{dom}(f) = [0, \infty)^d$ and $\text{int}(\text{dom}(f)) = (0, \infty)^d$. The gradient is $\nabla f(x) = -1/\sqrt{x}$, which blows up on sequence (x_n) approaching $\partial \text{int}(\text{dom}(f))$. Strict convexity is also obvious so f is Legendre. In Exercise 26.5 we ask you to calculate the Bregman divergences with respect to f and f^* and verify the results of Theorem 26.3.

The Taylor series of the Bregman divergence is often a useful approximation. Let $g(y) = D_f(x, y)$, which for $y = x$ has $\nabla g(y) = 0$ and $\nabla^2 g(y) = \nabla^2 f(x)$. A second order Taylor expansion suggests that

$$D_f(x, y) = g(y) \approx g(x) + \langle \nabla g(x), y - x \rangle + \frac{1}{2} \|y - x\|_{\nabla^2 f(x)}^2 = \frac{1}{2} \|y - x\|_{\nabla^2 f(x)}^2.$$

This approximation can be very poor if x and y are far apart. Even when x and y are close the lower order terms are occasionally problematic, but nevertheless the approximation can guide intuition. The next theorem, which is based on Taylor's theorem, gives an exact result.

THEOREM 26.4 *If f is convex and twice differentiable in $A = \text{int}(\text{dom}(f))$ and $x, y \in A$, then there exists an $\alpha \in [0, 1]$ and $z = \alpha x + (1 - \alpha)y$ such that*

$$D_f(x, y) = \frac{1}{2} \|x - y\|_{\nabla^2 f(z)}^2.$$

The next result will be useful.

THEOREM 26.5 *Let $\eta > 0$ and f be Legendre and twice differentiable in $A = \text{int}(\text{dom}(f))$. Let $z \in [x, y]$ be the point such that $D_f(x, y) = \frac{1}{2} \|x - y\|_{\nabla^2 f(z)}^2$. Then for all $u \in \mathbb{R}^d$,*

$$\langle x - y, u \rangle - \frac{D_f(x, y)}{\eta} \leq \frac{\eta}{2} \|u\|_{(\nabla^2 f(z))^{-1}}^2.$$

Proof Strict convexity of f ensures that $H = \nabla^2 f(z)$ is invertible. Applying Cauchy-Schwartz we have

$$\langle x - y, u \rangle \leq \|x - y\|_H \|u\|_{H^{-1}} = \|u\|_{H^{-1}} \sqrt{2D_f(x, y)}.$$

Therefore

$$\langle x - y, u \rangle - \frac{D_f(x, y)}{\eta} \leq \|u\|_{H^{-1}} \sqrt{2D_f(x, y)} - \frac{D_f(x, y)}{\eta} \leq \frac{\eta}{2} \|u\|_{H^{-1}}^2,$$

where the last step follows from the useful trick that $ax - bx^2 \leq a^2/(4b)$ for all $x \in \mathbb{R}$ and $b \geq 0$. \square

26.5 Optimization

The **first-order optimality condition** states that if $x \in \mathbb{R}^d$ is the minimizer of differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, then $\nabla f(x) = 0$. One of the things we

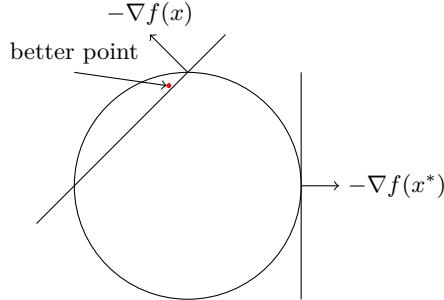


Figure 26.2 Illustration of first-order optimality conditions. The point at the top is not a minimizer because the hyperplane with normal as gradient does not support the convex set. The point at the right is a minimizer.

like about convex functions is that when f is convex the first-order optimality condition is both necessary and sufficient. The first-order optimality condition can also be generalized to constrained minima. In particular, if $A \subseteq \mathbb{R}^d$ is a nonempty convex set and $f : A \rightarrow \mathbb{R}$ is convex, then

$$x^* \in \operatorname{argmin}_{x \in A} f(x) \Leftrightarrow \forall x \in A : \langle \nabla f(x^*), x - x^* \rangle \geq 0. \quad (26.2)$$

The necessity of this condition is easy to understand by a geometric reasoning as shown in Fig. 26.2. Since x^* is a minimizer of f over A , $-\nabla f(x^*)$ must be the outer normal of a **supporting hyperplane** H_{x^*} of A at x^* otherwise x^* could be moved by a small amount while staying inside A and improving the value of f . Since A is convex, it thus lies entirely on the side of H_{x^*} that $\nabla f(x^*)$ points into. This is clearly equivalent to (26.2). The sufficiency of the condition also follows from this geometric viewpoint as the reader may verify from the figure.

The above statement continues to hold with a small modification even when f is not everywhere differentiable. In particular, in this case the equivalence (26.2) holds for any $x^* \in \operatorname{dom}(\nabla f)$ with the modification that on both sides of the equivalence, A should be replaced by $A \cap \operatorname{dom}(f)$:

PROPOSITION 26.1 *Let $f : \operatorname{dom}(f) \rightarrow \mathbb{R}$ be a convex function, $A \neq \emptyset$, $A \subset \mathbb{R}^d$ convex. Then for any $x^* \in \operatorname{dom}(\nabla f)$, it holds that:*

$$x^* \in \operatorname{argmin}_{x \in A \cap \operatorname{dom}(f)} f(a) \Leftrightarrow \forall x \in A \cap \operatorname{dom}(f) : \langle \nabla f(x^*), x - x^* \rangle \geq 0. \quad (26.3)$$

26.6 Projections

If $A \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$, then the Euclidean projection of x on A is $\Pi_A(x) = \operatorname{argmin}_{y \in A} \|x - y\|_2^2$. One can also project with respect to a Bregman divergence

induced by convex function f . Let $\Pi_{A,f}$ by

$$\Pi_{A,f}(x) = \operatorname{argmin}_{y \in A} D_f(y, x).$$

A property of the projection that will be exploited heavily in subsequent chapters is that minimizing a Legendre function f on a convex constrained set A is (usually) equivalent to finding the unconstrained minimum on the domain of f and then projecting that point onto A .

THEOREM 26.6 *Let f be Legendre, $A \subset \mathbb{R}^d$ a closed convex set and assume that $\tilde{y} = \operatorname{argmin}_{z \in \operatorname{dom}(f)} f(z)$ exists. Then the following hold:*

- (a) $y = \operatorname{argmin}_{z \in A \cap \operatorname{dom}(f)} f(z)$ exists and is unique;
- (b) $y = \operatorname{argmin}_{z \in A \cap \operatorname{dom}(f)} D_f(z, \tilde{y})$.

The assumption that \tilde{y} exists is necessary. For example $f(x) = -\sqrt{x}$ for $x \geq 0$ and $f(x) = \infty$ for $x < 0$ is Legendre with domain $\operatorname{dom}(f) = [0, \infty)$, but f does not have a minimum on its domain.

26.7 Notes

1 The ‘infinity arithmetic’ on the extended real line is as follows:

$$\begin{aligned}\alpha + \infty &= \infty && \text{for } \alpha \in (-\infty, \infty] \\ \alpha - \infty &= -\infty && \text{for } \alpha \in [-\infty, \infty) \\ \alpha \cdot \infty &= \infty \text{ and } \alpha \cdot (-\infty) = -\infty && \text{for } \alpha > 0 \\ \alpha \cdot \infty &= -\infty \text{ and } \alpha \cdot (-\infty) = \infty && \text{for } \alpha < 0 \\ 0 \cdot \infty &= 0 \cdot (-\infty) = 0.\end{aligned}$$

Like $\alpha/0$ the value of $\infty - \infty$ is not defined. We also have $\alpha \leq \infty$ for all α and $\alpha \geq -\infty$ for all α .

2 There are many ways to define the topological notions used in this chapter. The most elegant is also the most abstract, but we cannot do it justice here. Instead we give the classical definitions that are specific to \mathbb{R}^d and subsets. Let A be a subset of \mathbb{R}^d . A point $x \in A$ is an **interior point** if there exists an $\varepsilon > 0$ such that $B_\varepsilon(x) = \{y : \|x - y\|_2 \leq \varepsilon\} \subset A$. The **interior** of A is $\operatorname{int}(A) = \{x \in A : x \text{ is an interior point}\}$. The set A is **open** if $\operatorname{int}(A) = A$ and **closed** if its complement $A^c = \mathbb{R}^d \setminus A$ is open. The boundary of A is denoted by ∂A and is the set of points in $x \in \mathbb{R}^d$ such that for all $\varepsilon > 0$ the set $B_\varepsilon(x)$ contains points from A and A^c . Note that points in the boundary need not be in A . Some examples: $\partial \mathbb{R}^n = \emptyset$ and $\partial[0, \infty) = \{0\}$.

26.8 Bibliographic remarks

The main source for these notes is the excellent book by Rockafellar [2015]. The basic definitions are in Part I. The Fenchel dual is analyzed in Part III while Legendre functions are found in Part V. Convex optimization is a huge topic. The standard text is by Boyd and Vandenberghe [2004].

26.9 Exercises

26.1 For each of the real-valued functions below decide whether or not it is Legendre on the given domain.

- (a) $f(x) = x^2$ on $[-1, 1]$.
- (b) $f(x) = -\sqrt{x}$ on $[0, \infty)$.
- (c) $f(x) = \log(1/x)$ on $[0, \infty)$ with $f(0) = \infty$.
- (d) $f(x) = x \log(x)$ on $[0, \infty)$ with $f(0) = 0$.
- (e) $f(x) = |x|$ on \mathbb{R} .
- (f) $f(x) = \max\{|x|, x^2\}$ on \mathbb{R} .

26.2 Prove Theorem 26.2.

26.3 Prove Corollary 26.1.

26.4 Prove Proposition 26.1.

26.5 Let f be the convex function given in Example 26.4.

- (a) For $x, y \in \text{dom}(f)$ find $D_f(x, y)$.
- (b) Compute $f^*(u)$ and $\nabla f^*(u)$.
- (c) Find $\text{dom}(\nabla f^*)$.
- (d) Show that for $u, v \in (-\infty, 0]^d$,

$$D_{f^*}(u, v) = - \sum_{i=1}^d \frac{(u_i - v_i)^2}{u_i v_i^2}.$$

(e) Verify the claims in Theorem 26.3.

26.6 Let f be Legendre. Show that \tilde{f} given by $\tilde{f}(x) = f(x) + \langle x, u \rangle$ is also Legendre for any $u \in \mathbb{R}^d$.

26.7 Let f be the unnormalized negentropy function from Example 26.2.

- (a) Prove that f is Legendre.
- (b) Given $y \in [0, \infty)^d$, prove that $\text{argmin}_{x \in \mathcal{P}_{d-1}} D_f(x, y) = y/\|y\|_1$.

26.8 Let $\alpha \in [0, 1/d]$ and $\mathcal{A} = \mathcal{P}_{d-1} \cap [\alpha, 1]^d$ and f be the unnormalized negentropy function. Let $y \in [0, \infty)^d$ and $x = \operatorname{argmin}_{x \in \mathcal{A}} D_f(x, y)$ and assume that $y_1 \leq y_2 \leq \dots \leq y_d$. Let m be the smallest value such that $y_m / \sum_{i=m}^d y_i \geq \alpha$. Show that $x_i = \alpha$ if $i < m$ and $x_i = y_i / \sum_{j=m}^d y_j$ otherwise.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

27 Exp3 for Adversarial Linear Bandits

The model for adversarial linear bandits is as follows. The learner is given an action set $\mathcal{A} \subset \mathbb{R}^d$ and the number of rounds n . An instance of the adversarial problem is a sequence of loss vectors y_1, \dots, y_n where $y_t \in \mathbb{R}^d$ for each t . As usual in the adversarial setting, it is convenient to switch to losses. In each round t the learner selects an action $A_t \in \mathcal{A}$ and observes a loss $Y_t = \langle A_t, y_t \rangle$. The learner does not observe the loss vector y_t (if the loss vector is observed, then we call it the **full information setting**, but this is a topic for another book). Our standing assumption will be that the scalar loss for any of the action is in $[-1, 1]$, which corresponds to assuming that y_t is chosen from the polar of \mathcal{A} . For the rest of this chapter we assume that $y_t \in \mathcal{A}^\circ$ for all t . We furthermore assume that \mathcal{A} spans \mathbb{R}^d . The latter of these assumptions is for convenience only and may be relaxed with just a little care (Exercise 27.6). The regret of the learner after n rounds is

$$R_n = \mathbb{E} \left[\sum_{t=1}^n Y_t \right] - \min_{a \in \mathcal{A}} \sum_{t=1}^n \langle a, y_t \rangle.$$

Clearly the finite-armed adversarial bandits discussed in Chapter 11 is a special case of adversarial linear bandits corresponding to the choice $\mathcal{A} = \{e_1, \dots, e_d\}$ where e_1, \dots, e_d are the unit vectors of the d -dimensional standard Euclidean basis.

27.1 Exponential weights for linear bandits

We adapt the exponential weights algorithm of Chapter 11. Like in that setting we need a way to estimate the individual losses for each action, but now we make use of the linear structure to share information between the arms and decrease the variance of our estimators. For now we assume that \mathcal{A} is finite, which we relax in Section 27.3. Let $t \in [n]$ be the index of the current round. Assuming the loss estimate for action $a \in \mathcal{A}$ in round $s \in [n]$ is $\hat{Y}_s(a)$, then the distribution proposed by exponential weights is $P_t : \mathcal{A} \rightarrow [0, 1]$ given by

$$\tilde{P}_t(a) \propto \exp \left(-\eta \sum_{s=1}^{t-1} \hat{Y}_s(a) \right),$$

where $\eta > 0$ is the learning rate. To control the variance of the loss estimates, it will be useful to mix this distribution with an exploration distribution $\pi : \mathcal{A} \rightarrow [0, 1]$ with $\sum_{a \in \mathcal{A}} \pi(a) = 1$. The mixture distribution is

$$P_t(a) = (1 - \gamma)\tilde{P}_t(a) + \gamma\pi(a),$$

where γ is a constant mixing factor to be chosen later. The algorithm then simply samples its action A_t from P_t .

$$A_t \sim P_t.$$

Recall that $Y_t = \langle A_t, y_t \rangle$ is the observed loss after taking action A_t . We need a way to estimate $y_t(a) = \langle a, y_t \rangle$. The idea is to use least squares to estimate y_t with $\hat{Y}_t = R_t A_t Y_t$ where $R_t \in \mathbb{R}^{d \times d}$ is selected so that \hat{Y}_t is an unbiased estimate of y_t given the history. Then the loss for a given action is estimated by $\hat{Y}_t(a) = \langle a, \hat{Y}_t \rangle$. To find the choice of R_t that makes \hat{Y}_t unbiased let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | A_1, \dots, A_{t-1}]$ and calculate

$$\mathbb{E}_t[\hat{Y}_t] = R_t \mathbb{E}_t[A_t A_t^\top] y_t = R_t \underbrace{\left(\sum_a P_t(a) a a^\top \right)}_{Q_t} y_t.$$

Using $R_t = Q_t^{-1}$ leads to $\mathbb{E}_t[\hat{Y}_t] = y_t$ as desired. Of course Q_t should be non-singular, which will follow by choosing π so that

$$Q(\pi) = \sum_{a \in \mathcal{A}} \pi(a) a a^\top$$

is non-singular. The complete algorithm is summarized in Algorithm 14.

```

1: Input Action set  $\mathcal{A} \subset \mathbb{R}^d$ , learning rate  $\eta$ , exploration distribution  $\pi$ , exploration parameter  $\gamma$ 
2: for  $t = 1, 2, \dots, n$  do
3:   Compute sampling distribution:
```

$$P_t(a) = \gamma\pi(a) + (1 - \gamma) \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \hat{Y}_s(a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{Y}_s(a')\right)}.$$

```
4:   Sample action:
```

$$A_t \sim P_t.$$

```
5:   Observe loss  $Y_t = \langle A_t, y_t \rangle$  and compute loss estimates:
```

$$\hat{Y}_t = Q_t^{-1} A_t Y_t \quad \text{and} \quad \hat{Y}_t(a) = \langle a, \hat{Y}_t \rangle.$$

```
6: end for
```

Algorithm 14: Exp3 for Linear Bandits

27.2 Regret analysis

THEOREM 27.1 Assume that $\text{span}(\mathcal{A}) = \mathbb{R}^d$. There exists an exploration distribution π and parameters η and γ such that for all $(y_t)_t$ with $y_t \in \mathcal{A}^\circ$ the regret of Algorithm 14 is at most $R_n \leq 2\sqrt{3dn \log(K)}$.

Proof Assume that the learning rate η is chosen so that for each round t the loss estimates satisfy

$$\eta \hat{Y}_t(a) \geq -1, \quad \forall a \in \mathcal{A}. \quad (27.1)$$

Then by modifying the proof of Theorem 11.1 (see Exercise 27.1) the regret is bounded by

$$R_n \leq \frac{\log K}{\eta} + 2\gamma n + \eta \sum_t \mathbb{E} \left[\sum_a P_t(a) \hat{Y}_t^2(a) \right]. \quad (27.2)$$

Note that we cannot use the proof that leads to the tighter constant (η getting replaced by $\eta/2$ in the second term above) because there is no guarantee that the loss estimates will be upper bounded by one. To get a regret bound it remains to set γ and η so that (27.1) is satisfied and to bound $\mathbb{E} \left[\sum_a P_t(a) \hat{Y}_t^2(a) \right]$. We start with the latter. Let $M_t = \sum_a P_t(a) \hat{Y}_t^2(a)$. By definition of the loss estimate,

$$\hat{Y}_t^2(a) = (a^\top Q_t^{-1} A_t Y_t)^2 = Y_t^2 A_t^\top Q_t^{-1} a a^\top Q_t^{-1} A_t,$$

which means that $M_t = \sum_a P_t(a) \hat{Y}_t^2(a) = Y_t^2 A_t^\top Q_t^{-1} A_t \leq A_t^\top Q_t^{-1} A_t$ and

$$\mathbb{E}_t[M_t] \leq \text{trace} \left(\sum_a P_t(a) a a^\top Q_t^{-1} \right) = d.$$

It remains to choose γ and η . Strengthen (27.1) to $|\eta \hat{Y}_t(a)| \leq 1$ and note that since $|Y_t| \leq 1$,

$$|\eta \hat{Y}_t(a)| = |\eta a^\top Q_t^{-1} A_t Y_t| \leq \eta |a^\top Q_t^{-1} A_t|.$$

Let $Q(\pi) = \sum_{\nu \in \mathcal{A}} \pi(\nu) \nu \nu^\top$. Clearly $Q_t \succeq \gamma Q(\pi)$ and hence $Q_t^{-1} \preceq Q(\pi)^{-1}/\gamma$ by Exercise 27.3. Using this and Cauchy-Schwartz inequality shows that

$$|a^\top Q_t^{-1} A_t| \leq \|a\|_{Q_t^{-1}} \|A_t\|_{Q_t^{-1}} \leq \max_{\nu \in \mathcal{A}} \nu^\top Q_t^{-1} \nu \leq \frac{1}{\gamma} \max_{\nu \in \mathcal{A}} \nu^\top Q^{-1}(\pi) \nu,$$

which implies that

$$|\eta \hat{Y}_t(a)| \leq \frac{\eta}{\gamma} \max_{\nu \in \mathcal{A}} \nu^\top Q^{-1}(\pi) \nu = \frac{\eta}{\gamma} \max_{\nu \in \mathcal{A}} \|\nu\|_{Q^{-1}(\pi)}^2. \quad (27.3)$$

From Theorem 21.1 (Kiefer–Wolfowitz) we know there exists a sampling distribution π such that $\max_{\nu \in \mathcal{A}} \|\nu\|_{Q^{-1}(\pi)}^2 = d$. By choosing $\gamma = \eta d$ and plugging into (27.2) we get

$$R_n \leq \frac{\log K}{\eta} + 3\eta dn = 2\sqrt{3dn \log(K)},$$

where the last equality is derived by choosing $\eta = \sqrt{\frac{\log(K)}{3dn}}$. \square

27.3 Continuous exponential weights

As the number of arms becomes extremely large or infinite, the dependence on $\log(K)$ may be undesirable. Suppose that $\mathcal{A} \subset [-1, 1]^d$ is a subset of the hypercube and K is extremely large. Letting $\varepsilon = 1/n$ define $\mathcal{A}' \subset \mathcal{A}$ to be the smallest set such that for all $x \in \mathcal{A}$ there exists a $y \in \mathcal{A}'$ with $|\langle x - y, u \rangle| \leq \varepsilon$ for all $u \in \mathcal{A}^\circ$. That is, \mathcal{A}' is an ε -accurate approximation (loss-wise) to \mathcal{A} . A standard calculation (cf. Exercise 27.5) shows that no matter how large is K , the set \mathcal{A}' is guaranteed to satisfy $\log |\mathcal{A}'| = O(d \log n)$. Then it is easy to check that playing Exp3 on \mathcal{A}' suffers regret at most $R_n = O(d\sqrt{n \log(n)})$. Notice that this even works when $|\mathcal{A}| = \infty$. The problem with this approach is that \mathcal{A}' is still exponentially large, which makes the running time of Exp3 unreasonably costly. When \mathcal{A} is itself a convex set, then a more computationally tractable approach is to switch to the **continuous exponential weights** algorithm.

Let $\pi : \mathcal{A} \rightarrow [0, 1]$ be the same exploration distribution as used in the proof of the previous section. The continuous exponential weights policy samples A_t from $P_t = (1 - \gamma)\tilde{P}_t + \gamma\pi(a)$ where \tilde{P}_t a measure supported on \mathcal{A} defined by

$$\tilde{P}_t(A) = \frac{\int_A \exp\left(-\eta \sum_{s=1}^{t-1} \hat{Y}_s(a)\right) da}{\int_{\mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{Y}_s(a)\right) da}. \quad (27.4)$$

We will shortly see that the analysis in the previous section can be copied almost verbatim to prove a regret bound for this strategy. But what has been bought here? Rather than sampling from a discrete distribution on a large number of arms we now have to sample from a continuous density on a convex set. Sampling from arbitrary densities is itself a challenging problem, but under certain conditions there are polynomial time algorithms for this problem. The factors that play the biggest role in the feasibility of sampling from a distribution are (a) what is the form of the distribution and (b) how is the convex set represented. As it happens the measure defined in the last display is **log-concave**, which means that the logarithm of the density is a concave function (in this case it is even a linear function). Suppose that $p_t(a) \propto \mathbb{1}_{\mathcal{A}}(a) \exp(-f(a))$ is a density with respect to the Lebesgue measure on \mathcal{A} , then there exists a polynomial-time algorithm for sampling from p provided one can compute:

- 1 (first-order information): $\nabla f(a)$ and for any $a \in \mathcal{A}$.
- 2 (projections): For all $y \in \mathbb{R}^d$ find $x = \operatorname{argmin}_{x \in \mathcal{A}} \|x - y\|_2$.

Clearly for densities defined by Eq. (27.4) satisfy the first condition. Efficiently computing a projection onto a convex set is a more delicate issue. A general criteria that makes this efficient is access to a **separation oracle**, which is a pair of functions $\phi : \mathbb{R}^d \rightarrow \{0, 1\}$ and $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $\phi(y) = \mathbb{1}_{\mathcal{A}}(y)$ and

$\psi(y)$ is arbitrary for $y \in \mathcal{A}$ and for $y \notin \mathcal{A}$, $u = \psi(y)$ satisfies $\langle x - y, u \rangle \leq 0$ for all $x \in \mathcal{A}$. That is, the separation oracle accepts points in \mathbb{R}^d as input and responds as output whether or not that point is inside the set and if it is not provides a separating hyperplane.

The analysis of the exponential weights algorithm goes through almost unchanged. By repeating the analysis in the previous section, but replacing sums with integrals one obtains the following bound on the regret.

THEOREM 27.2 *The regret of continuous exponential weights algorithm is bounded by*

$$R_n \leq \frac{1}{\eta} \log \left(\frac{\text{vol}(\mathcal{A})}{\int_{\mathcal{A}} \exp \left(-\eta \sum_{t=1}^n (\hat{Y}_t(a) - \hat{Y}_t(a^*)) \right) da} \right) + \gamma n + \eta dn, \quad (27.5)$$

where $\text{vol}(\mathcal{A}) = \int_{\mathcal{A}} da$ is the volume of the action set \mathcal{A} .

The term inside the logarithm is bounded using the following proposition, the proof of which we leave as an exercise to the reader.

PROPOSITION 27.1 *Let $\mathcal{K} \subset \mathbb{R}^d$ be a compact convex set with $\text{vol}(\mathcal{K}) > 0$ and $u \in \mathbb{R}^d$ and $x^* = \operatorname{argmin}_{x \in \mathcal{K}} \langle x, u \rangle$. Then*

$$\log \left(\frac{\text{vol}(\mathcal{K})}{\int_{\mathcal{K}} \exp(-\langle x^* - x, u \rangle) dx} \right) \leq 1 + \max \left(0, d \log \left(\sup_{x, y \in \mathcal{K}} \langle x - y, u \rangle \right) \right).$$

Substituting this result into Eq. (27.5) and choosing $\eta = \sqrt{\log(n)/n}$ leads to

$$R_n = O(d\sqrt{n \log(n)}),$$

which matches the bound we got from the discretization approach.

27.4 Notes

- 1 A naive implementation of Algorithm 14 has computation complexity $O(Kd + d^3)$ per round. There is also the one-off cost of computing the exploration distribution, the complexity of which is discussed in Chapter 21. The real problem is that K can be extremely large. This is especially true when the action set is combinatorial. For example, when $\mathcal{A} = \{a \in \mathbb{R}^d : a_i = \pm 1\}$ is the corners of the hypercube $|\mathcal{A}| = 2^d$, which is much too large unless the dimension is small. Such problems call for a different approach that we present in the next chapter and in Chapter 30.
- 2 It is not important to find exactly the optimal exploration distribution. All that is needed is a bound on Eq. (27.3), which for the exploration distribution based on John's ellipsoid is just d .
- 3 We will discuss lower bounds in Chapter 29. In general neither Algorithm 14 or its continuous variant lead to near-optimal regret in the minimax sense on

some action sets. An example where this occurs when $\mathcal{A} = \mathcal{A}^\circ$ are the unit ball, which is a topic for the next chapter.

- 4 Like for stochastic bandits, Algorithm 14 and Theorem 27.1 can be generalized to the case when the action set is different in each round. The only adjustment to the algorithm is that now the exploration distribution must be recomputed in every round. The analysis goes through without change.

27.5 Bibliographic remarks

The results in Sections 27.1 and 27.2 follow the article by Bubeck et al. [2012] with minor modifications to make the argument more pedagogical. The main difference is that they used John's ellipsoid over the action set for exploration, which is only the 'right thing' when the John's ellipsoid is also a central ellipsoid. Here we use Kiefer–Wolfowitz, which is equivalent to finding the minimum volume central ellipsoid containing the action set as described in Chapter 21, where we also discuss the computation properties of finding the core set necessary to define the exploration distribution. A polynomial time sampling algorithm for convex sets with gradient information and projections is by Bubeck et al. [2015b]. We warn the reader that these algorithms are perhaps not the most practical, especially if theoretically justified parameters are used. The study of sampling from convex bodies is quite fascinating. There is a overview by Lovász and Vempala [2007], though it is a little old. Another path towards an efficient $O(d\sqrt{n \log(\cdot)})$ policy for convex action sets is to use the tools from online optimization. We explain these ideas in more detail in the next chapter, but the reader is referred to the paper by Bubeck and Eldan [2015]. The continuous exponential weights algorithm is perhaps attributable to Cover [1991] in the special setting of online learning called universal portfolio optimization. The first application to linear bandits is by Hazan et al. [2016]. Their algorithm and analysis is more complicated because they seek to improve the computation properties by replacing John's exploration with an adaptive randomized exploration basis that under quite weak assumptions can be computed in polynomial time. Continuous exponential weights with John's exploration was recently analyzed by van der Hoeven et al. [2018].

27.6 Exercises

27.1 Prove Eq. (27.2).

27.2 Suppose that instead of assuming $y_t \in \mathcal{A}^\circ$ we assume that $y_t \in \{y \in \mathbb{R}^d : \sup_{a \in \mathcal{A}} |\langle a, y \rangle| \leq b\}$ for some known $b > 0$. Modify the algorithm to accommodate this change and explain how the regret guarantee changes.

27.3 Let $A, B \in \mathbb{R}^{d \times d}$ and suppose that $A \succeq B$ and B is invertible. Show that $A^{-1} \preceq B^{-1}$.

27.4 Now suppose that $a < b$ are known and $y_t \in \{y \in \mathbb{R}^d : \langle a, y \rangle \in [a, b] \text{ for all } a \in \mathcal{A}\}$. How can you adapt the algorithm now and what is its regret?

27.5 Let $\mathcal{A} \subset \mathbb{R}^d$ be bounded, $\|x\| = \sup_{u \in \mathcal{A}^\circ} |\langle x, u \rangle|$, where we also allow $\|x\| = \infty$. For $\mathcal{A}' \subset \mathcal{A}$ let $d(\mathcal{A}', \mathcal{A}) = \sup_{x \in \mathcal{A}} \inf_{y \in \mathcal{A}'} \|x - y\|$. Finally, for $\varepsilon > 0$, we let $N(\varepsilon, \mathcal{A})$ be the ε -covering number of \mathcal{A} , which is defined as in Definition 20.2 except that the Euclidean norm $\|\cdot\|_2$ used there is replaced by $\|\cdot\|$ defined above. Show that the following hold:

- (a) $\|\cdot\|$ satisfies the triangle inequality, $\|0\| = 0$ and $\|cx\| = |c|\|x\|$ for any $x \in \mathbb{R}^d$ and $c \in \mathbb{R}$;
- (b) Let $\mathcal{A}' \subset \mathcal{A}$ be finite. Show that $d(\mathcal{A}', \mathcal{A}) \leq \varepsilon$ if and only if \mathcal{A}' is an ε -cover of \mathcal{A} ;
- (c) Let $B \doteq \{z : \|z\| \leq 1\}$ and let $\mathcal{A}^* = \mathcal{A} \cup -\mathcal{A}$ denote the reflection of \mathcal{A} in the origin. Then, $\text{co}(\mathcal{A}^*) \subset B \subset \text{span}(\mathcal{A})$ and $\{z \in \mathbb{R}^d : \|z\| < \infty\} = \text{span}(\mathcal{A})$;
- (d) Let $p = \dim(\text{span}(\mathcal{A}))$. There exists a constant $c > 0$ that depends on \mathcal{A} such that for any $\varepsilon \leq 1/2$ the inequality $\log N(\varepsilon, \mathcal{A}) \leq cp \log(1/\varepsilon)$ holds;
- (e) For any $\varepsilon \leq 1/2$ there exists \mathcal{A}' such that $d(\mathcal{A}', \mathcal{A}) \leq \varepsilon$ and $\log |\mathcal{A}'| \leq cp \log(1/\varepsilon) \leq cd \log(1/\varepsilon)$.

(Hint: You should be aware of Exercise 20.1.)



One can also show that there exists some constant $c > 0$ such that for all $x \in \text{span}(\mathcal{A})$, $\|x\| \leq c\|x\|_2$. This implies that $\|\cdot\|$, when restricted to $\text{span}(\mathcal{A})$, is a norm.

27.6 In the definition of the algorithm and the proof of Theorem 27.1 we assumed that \mathcal{A} spans \mathbb{R}^d . Show that this assumption may be relaxed by carefully adapting the algorithm and analysis.

27.7 We saw in Chapter 11 that the exponential weights algorithm achieved near-optimal regret without mixing additional exploration. Show that exploration is crucial here. More precisely, construct a finite action set \mathcal{A} and reward sequence $y_t \in \mathcal{A}^\circ$ such that the regret of Algorithm 14 with $\gamma = 0$ leads to a very bad algorithm relative to the optimal choice.

27.8 Prove Theorem 27.2.

27.9 Prove Proposition 27.1.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

28 Follow the Regularized Leader and Mirror Descent

In the last chapter we showed that if $\mathcal{A} \subset \mathbb{R}^d$ has K elements, then the regret of Exp3 combined with John's exploration has regret

$$R_n = O(\sqrt{dn \log(K)}).$$

When \mathcal{A} is a convex set we also showed the continuous version of this algorithm has regret at most

$$R_n = O(d\sqrt{n \log(n)}).$$

Although this algorithm runs in polynomial time, the degree is high and the implementation is complicated and impractical. In many cases this can be improved upon, both in terms of the regret and computation. One of the main results of this chapter is a proof that when \mathcal{A} is the unit ball, then there is an efficient algorithm for which the regret is $R_n = O(\sqrt{dn \log(n)})$. More importantly, however, we introduce a pair of related algorithms called **follow the regularized leader** and **mirror descent**, which have proven to be powerful and flexible tools for the design and analysis of bandit algorithms.

28.1 Online linear optimization

Mirror descent originated in the convex optimization literature. The idea has since been adapted to online learning and specifically to online linear optimization. Online linear optimization is the full information version of the adversarial linear bandit where at the end of each round the learner observes the full vector y_t . Let $\mathcal{A} \subset \mathbb{R}^d$ be a convex set and $\mathcal{L} \subset \mathbb{R}^d$ be an arbitrary set called the **loss space**. Let y_1, \dots, y_n be a sequence of loss vectors with $y_t \in \mathcal{L}$ for all $t \in [n]$. In each round the learner chooses $a_t \in \mathcal{A}$ and subsequently observes y_t . The regret relative to a fixed comparator $a \in \mathcal{A}$ is

$$R_n(a) = \sum_{t=1}^n \langle a_t - a, y_t \rangle$$

and the regret is $R_n = \max_{a \in \mathcal{A}} R_n(a)$. We emphasize that the only difference relative to the adversarial linear bandit is that now y_t is observed rather than $\langle a_t, y_t \rangle$. Actions are not capitalized in this section because the algorithms presented here do not randomize.

Mirror descent

The basic version of mirror descent has two extra parameters beyond n and \mathcal{A} . A learning rate $\eta > 0$ and a Legendre function $F : \mathbb{R}^d \rightarrow \mathbb{R}$ with domain $\mathcal{D} = \text{dom}(F)$. The function F is usually called a **potential function** or **regularizer**. In the first round mirror descent predicts

$$a_1 = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} F(a), \quad (28.1)$$

In subsequent rounds it predicts

$$a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} (\eta \langle a, y_t \rangle + D_F(a, a_t)), \quad (28.2)$$

where $D_F(a, a_t)$ is the F -induced Bregman divergence between a and a_t . The minimization in (28.2) is over $\mathcal{A} \cap \mathcal{D}$ because for $a_t \in \text{int}(\mathcal{D}) \subseteq \text{dom}(\nabla F)$ the domain of $D_F(\cdot, a_t)$ is the same as that of F .

Follow the regularized leader

Like mirror descent, follow the regularized leader depends on a convex potential F with domain $\mathcal{D} = \text{dom}(F)$ and predicts $a_1 = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} F(a)$. In subsequent rounds it predicts

$$a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} \left(\eta \sum_{s=1}^t \langle a, y_s \rangle + F(a) \right). \quad (28.3)$$

The intuition is that the algorithm chooses a_{t+1} to be the action that performed best in hindsight with respect to the regularized loss. As for mirror descent, the regularization serves to stabilize the algorithm, which turns out to be a key property of good algorithms for online linear prediction.

 Another algorithm is **follow the leader**, which chooses the action that appears best in hindsight, $a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A}} \sum_{s=1}^t \langle a, y_s \rangle$. In general, this algorithm is not well suited for online linear optimization because the absence of regularization makes for an unstable algorithm that can lead to extremely poor performance as you will show in Exercise 28.2.

Equivalence of mirror descent and follow the regularized leader

At first sight these algorithms do not look that similar. To clarify matters let us suppose that F has domain $\mathcal{D} \subseteq \mathcal{A}$. We now show that in this setting mirror descent and follow the regularized leader are identical. Let

$$\Phi_t(a) = \eta \langle a, y_t \rangle + D_F(a, a_t) = \eta \langle a, y_t \rangle + F(a) - F(a_t) - \langle \nabla F(a_t), a - a_t \rangle.$$

Now mirror descent chooses a_{t+1} to minimize Φ_t . The reader should check that the assumption that F is Legendre on domain $\mathcal{D} \subseteq \mathcal{A}$ implies that the minimizer

occurs in the interior of $\mathcal{D} \subseteq \mathcal{A}$ and that $\nabla\Phi_t(a_{t+1}) = 0$ (cf. Exercise 28.1). This means that $\eta y_t = \nabla F(a_t) - \nabla F(a_{t+1})$ and so

$$\nabla F(a_{t+1}) = -\eta y_t + \nabla F(a_t) = \nabla F(a_1) - \eta \sum_{s=1}^t y_s = -\eta \sum_{s=1}^t y_s,$$

where the last equality is true because a_1 is chosen as the minimizer of F in $\mathcal{A} \cap \mathcal{D} = \mathcal{D}$ and again the fact that F is Legendre ensures this minimum occurs at an interior point where the gradient vanishes. Follow the regularized leader chooses a_{t+1} to minimize $\Phi'_t(a) = \eta \sum_{s=1}^t \langle a, y_s \rangle + F(a)$. The same argument shows that $\nabla\Phi'_t(a_{t+1}) = 0$, which means that

$$\nabla F(a_{t+1}) = -\eta \sum_{s=1}^t y_s.$$

The last two displays and the fact that the gradient for Legendre functions is invertible shows that mirror descent and follow the regularized leader are the same in this setting.



The equivalence between these algorithms is far from universal. First of all, it does not hold when F is not Legendre or its domain is larger than \mathcal{A} . Second, in many applications of these algorithms the learning rate or potential change with time and in either case the algorithms will typically produce different action sequences. For example, if a learning rate η_t is used rather than η in the definition of Φ_t , then mirror descent chooses $\nabla F(a_{t+1}) = -\sum_{s=1}^t \eta_s y_s$ while follow the regularized leader chooses $\nabla F(a_{t+1}) = -\eta_t \sum_{s=1}^t y_s$. We return to this issue in the notes and exercises.

EXAMPLE 28.1 Let $\mathcal{A} = \mathbb{R}^d$ and $F(a) = \frac{1}{2}\|a\|_2^2$. Then $\nabla F(a) = a$ and $D(a, a_t) = \frac{1}{2}\|a - a_t\|_2^2$. Clearly F is Legendre and $\mathcal{D} = \mathcal{A}$ so mirror descent and follow the regularized leader are the same. By simple calculus we see that

$$a_{t+1} = \operatorname{argmin}_{a \in \mathbb{R}^d} \eta \langle a, y_t \rangle + \frac{1}{2} \|a - a_t\|_2^2 = a_t - \eta y_t,$$

which may be familiar as **online gradient descent**.

EXAMPLE 28.2 Let \mathcal{A} be a compact subset of \mathbb{R}^d and $F(a) = \frac{1}{2}\|a\|_2^2$. Then mirror descent chooses

$$a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A}} \eta \langle a, y_t \rangle + \frac{1}{2} \|a - a_t\|_2^2 = \Pi(a_t - \eta y_t), \quad (28.4)$$

where $\Pi(a)$ is the Euclidean projection of a onto \mathcal{A} . This algorithm is usually called online projected gradient descent. On the other hand, for follow the regularized leader we have

$$a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A}} \eta \sum_{s=1}^t \langle a, y_s \rangle + \frac{1}{2} \|a - a_t\|_2^2 = \Pi \left(-\eta \sum_{s=1}^t y_s \right),$$

which may be a different choice than mirror descent.

EXAMPLE 28.3 The exponential weights algorithm that appeared on numerous occasions in earlier chapters is a special case of mirror descent corresponding to choosing the constraint set \mathcal{A} as the simplex in \mathbb{R}^d and choosing F to be the unnormalized negentropy function of Example 26.2.

A two-step process for implementation

Solving the optimization problem in Eq. (28.2) is often made easier by using the results in Section 26.6 of Chapter 26. Let $\mathcal{D}^* = \nabla F(\mathcal{D})$ and suppose the following condition holds:

$$\nabla F(x) - \eta y \in \mathcal{D}^* \text{ for all } x \in \mathcal{A} \cap \mathcal{D} \text{ and } y \in \mathcal{L}. \quad (28.5)$$

Then solution to Eq. (28.2) can be found using the following two-step procedure.

$$\tilde{a}_{t+1} = \operatorname{argmin}_{a \in \mathcal{D}} \eta \langle a, y_t \rangle + D_F(a, a_t) \quad \text{and} \quad (28.6)$$

$$a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} D_F(a, \tilde{a}_{t+1}). \quad (28.7)$$

Eq. (28.5) means the first optimization problem can be evaluated explicitly as the solution to

$$\eta y_t + \nabla F(\tilde{a}_{t+1}) - \nabla F(a_t) = 0. \quad (28.8)$$

Since F is Legendre this means that $\tilde{a}_{t+1} = (\nabla F)^{-1}(\nabla F(a_t) - \eta y_t)$. The optimization problem in Eq. (28.7) is usually harder to calculate analytically, but there are important exceptions as we shall see.



The condition in Eq. (28.5) holds for all choices of potential and losses in this book.

28.2 Regret analysis

We now analyze the regret of mirror descent. The theorem has two parts, the first of which is strictly stronger by a small margin than the second. To minimize clutter we abbreviate D_F by D .

THEOREM 28.1 *Let $\eta > 0$ and F be Legendre with domain \mathcal{D} and $\mathcal{A} \subseteq \operatorname{cl}(\mathcal{D})$. Let a_1, \dots, a_{n+1} be the actions chosen by mirror descent. Then for any $a \in \mathcal{A}$ the regret of mirror descent is bounded by:*

$$R_n(a) \leq \sum_{t=1}^n \langle a_t - a_{t+1}, y_t \rangle + \frac{F(a) - F(a_1)}{\eta} - \frac{1}{\eta} \sum_{t=1}^n D(a_{t+1}, a_t).$$

Furthermore, suppose that Eq. (28.5) holds and $\tilde{a}_2, \tilde{a}_3, \dots, \tilde{a}_{n+1}$ are given by Eq. (28.6), then

$$R_n(a) \leq \frac{1}{\eta} \left(F(a) - F(a_1) + \sum_{t=1}^n D(a_t, \tilde{a}_{t+1}) \right).$$

Proof of Theorem 28.1 For the first part we split the inner product:

$$\langle a_t - a, y_t \rangle = \langle a_t - a_{t+1}, y_t \rangle + \langle a_{t+1} - a, y_t \rangle.$$

Using the fact that $a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} \eta \langle a, y_t \rangle + D(a, a_t)$ and the first order optimality conditions shows that

$$\langle \eta y_t + \nabla F(a) - \nabla F(a_t), a - a_{t+1} \rangle \geq 0.$$

By the definition of the Bregman divergence we have

$$\begin{aligned} \langle a_{t+1} - a, y_t \rangle &\leq \frac{1}{\eta} \langle \nabla F(a) - \nabla F(a_t), a - a_{t+1} \rangle \\ &= \frac{1}{\eta} (D(a, a_t) - D(a, a_{t+1}) - D(a_{t+1}, a_t)). \end{aligned} \quad (28.9)$$

Using this, along with the definition of the regret,

$$\begin{aligned} R_n &= \sum_{t=1}^n \langle a_t - a, y_t \rangle \\ &\leq \sum_{t=1}^n \langle a_t - a_{t+1}, y_t \rangle + \frac{1}{\eta} \sum_{t=1}^n (D(a, a_t) - D(a, a_{t+1}) - D(a_{t+1}, a_t)) \\ &= \sum_{t=1}^n \langle a_t - a_{t+1}, y_t \rangle + \frac{1}{\eta} \left(D(a, a_1) - D(a, a_{n+1}) - \sum_{t=1}^n D(a_{t+1}, a_t) \right) \\ &\leq \sum_{t=1}^n \langle a_t - a_{t+1}, y_t \rangle + \frac{F(a) - F(a_1)}{\eta} - \frac{1}{\eta} \sum_{t=1}^n D(a_{t+1}, a_t), \end{aligned}$$

where the final inequality follows from the fact that $D(a, a_{n+1}) \geq 0$ and $D(a, a_1) \leq F(a) - F(a_1)$, which is true by the first-order optimality conditions for $a_1 = \operatorname{argmin}_{b \in \mathcal{A}} F(b)$. To see the second part note that

$$\begin{aligned} \langle a_t - a_{t+1}, y_t \rangle &= \frac{1}{\eta} \langle a_t - a_{t+1}, \nabla F(a_t) - \nabla F(\tilde{a}_{t+1}) \rangle \\ &= \frac{1}{\eta} (D(a_{t+1}, a_t) + D(a_t, \tilde{a}_{t+1}) - D(a_{t+1}, \tilde{a}_{t+1})) \\ &\leq \frac{1}{\eta} (D(a_{t+1}, a_t) + D(a_t, \tilde{a}_{t+1})). \end{aligned}$$

The result follows by substituting this into Eq. (28.9) and completing as for the first part. \square



The assumption that a_1 minimizes the potential was only used to bound $D(a, a_1) \leq F(a) - F(a_1)$. For a different initialization the following bound still holds:

$$R_n(a) \leq \frac{1}{\eta} \left(D(a, a_1) + \sum_{t=1}^n D(a_t, \tilde{a}_{t+1}) \right).$$

As we shall see in Chapter 31, this is useful when using mirror descent to analyze nonstationary bandits.

The first part of Theorem 28.1 also holds for follow the regularized leader as stated in the next result, the proof of which is left for Exercise 28.6.

THEOREM 28.2 *Let $\eta > 0$ and F be Legendre with domain \mathcal{D} and $\mathcal{A} \subseteq \text{cl}(\mathcal{D})$. Then for any $a \in \mathcal{A}$ the regret of follow the regularized leader is bounded by*

$$R_n(a) \leq \sum_{t=1}^n \langle a_t - a_{t+1}, y_t \rangle + \frac{F(a) - F(a_1)}{\eta} - \frac{1}{\eta} \sum_{t=1}^n D(a_{t+1}, a_t).$$

We now give two examples of how to apply Theorem 28.1. Let $\text{diam}_F(\mathcal{A}) = \max_{a,b \in \mathcal{A}} |F(a) - F(b)|$ be the diameter of \mathcal{A} with respect to F .

PROPOSITION 28.1 *Let $\mathcal{A} = B_2^d = \{a \in \mathbb{R}^d : \|a\|_2 \leq 1\}$ be the standard unit ball and assume $y_t \in B_2^d$ for all t . Then mirror descent with potential $F(a) = \frac{1}{2}\|a\|_2^2$ and $\eta = \sqrt{1/n}$ satisfies $R_n \leq \sqrt{n}$.*

Proof By Eq. (28.8) we have $\tilde{a}_{t+1} = a_t - \eta y_t$ so

$$D(a_t, \tilde{a}_{t+1}) = \frac{1}{2} \|\tilde{a}_{t+1} - a_t\|_2^2 = \frac{\eta^2}{2} \|y_t\|_2^2.$$

Therefore since $\text{diam}_F(\mathcal{A}) = 1/2$ and $\|y_t\|_2 \leq 1$ for all t ,

$$R_n \leq \frac{\text{diam}_F(\mathcal{A})}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \|y_t\|_2^2 \leq \frac{1}{2\eta} + \frac{\eta n}{2} = \sqrt{n}. \quad \square$$

PROPOSITION 28.2 *Let $\mathcal{A} = \{a \in [0, 1]^d : \sum_{i=1}^d a_i = 1\}$ be the probability simplex and $y_t \in \mathcal{A}^\circ$ for all t . Then mirror descent with the unnormalized negentropy potential and $\eta = \sqrt{2 \log(d)/n}$ satisfies $R_n \leq \sqrt{2n \log(d)}$.*

Proof The Bregman divergence with respect to the unnormalized negentropy

potential for $a, b \in \mathcal{A}$ is $D(a, b) = \sum_{i=1}^d a_i \log(a_i/b_i)$. Therefore

$$\begin{aligned} R_n(a) &\leq \frac{F(a) - F(a_1)}{\eta} + \sum_{t=1}^n \langle a_t - a_{t+1}, y_t \rangle - \frac{1}{\eta} \sum_{t=1}^n D(a_{t+1}, a_t) \\ &\leq \frac{\log(d)}{\eta} + \sum_{t=1}^n \|a_t - a_{t+1}\|_1 \|y_t\|_\infty - \frac{1}{\eta} \sum_{t=1}^n \frac{1}{2} \|a_t - a_{t+1}\|_1^2 \\ &\leq \frac{\log(d)}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \|y_t\|_\infty^2 \leq \frac{\log(d)}{\eta} + \frac{\eta n}{2} = \sqrt{2n \log(d)}. \end{aligned}$$

where the first inequality follows from Theorem 28.1, the second from Pinsker's inequality and the facts that $\text{diam}_F(\mathcal{A}) = \log(d)$. In the third inequality we used that fact that $ax - bx^2/2 \leq a^2/(2b)$ for all x . The last inequality follows from the assumption that $\|y_t\|_\infty \leq 1$. \square

The last few steps in the above proof are so routine that we summarize their use in a corollary, the proof of which we leave to the reader (Exercise 28.3).

COROLLARY 28.1 *Let F be a Legendre potential and $\|\cdot\|_t$ be a norm on \mathbb{R}^d for each $t \in [n]$ such that $D_F(a_{t+1}, a_t) \geq \frac{1}{2} \|a_{t+1} - a_t\|_t^2$. Then the regret of mirror descent or follow the regularized leader satisfies*

$$R_n \leq \frac{\text{diam}_F(\mathcal{A})}{\eta} + \frac{\eta}{2} \sum_{t=1}^n (\|y_t\|_t^*)^2,$$

where $\|y\|_t^* = \max_{x: \|x\|_t \leq 1} \langle x, y \rangle$ is the **dual norm** of $\|\cdot\|_t$.

It often happens that the easiest way to bound the regret of mirror descent is to find a norm that satisfies the conditions of Corollary 28.1. To illustrate a suboptimal application of mirror descent and this result, suppose we had chosen $F(a) = \frac{1}{2} \|a\|_2^2$ in the setting of Proposition 28.2. Then $D_F(a_{t+1}, a_t) = \frac{1}{2} \|a_{t+1} - a_t\|^2$ suggests choosing $\|\cdot\|_t$ to be the standard Euclidean norm. Since $\text{diam}_F(\mathcal{A}) = 1/2$ and $\|\cdot\|_2^* = \|\cdot\|_2$, applying Corollary 28.1 shows that

$$R_n \leq \frac{1}{2\eta} + \frac{\eta}{2} \sum_{t=1}^n \|y_t\|_2^2.$$

But now we see that $\|y_t\|_2^2$ can be as large as d and tuning η would lead to a rate of $O(\sqrt{nd})$ rather than $O(\sqrt{n \log(d)})$.

 Both theorems were presented for the oblivious case where (y_t) are chosen in advance. This assumption was not used, however, and in fact the bounds in this section continue to hold when y_t are chosen strategically as a function of $y_1, x_1, \dots, y_{t-1}, x_t$. This is analogous to how the basic regret bound for exponential weights continues to hold in the face of strategic losses. But be cautioned, this result does not carry immediately to the application of mirror descent to bandits as discussed at the end in Note 7.

28.3 Online learning for bandits

We now consider the application of mirror descent to bandit problems. Like in the previous chapter the adversary chooses a sequence of vectors y_1, \dots, y_n with $y_t \in \mathcal{L} \subset \mathbb{R}^d$. In each round the learner chooses $A_t \in \mathcal{A} \subset \mathbb{R}^d$ where \mathcal{A} is convex and observes $\langle A_t, y_t \rangle$. The regret relative to action a is

$$R_n(a) = \mathbb{E} \left[\sum_{t=1}^n \langle A_t - a, y_t \rangle \right].$$

The regret is $R_n = \max_{a \in \mathcal{A}} R_n(a)$. The application of mirror descent and follow the regularized leader to linear bandits requires two new ideas. First, because the learner only observes $\langle A_t, y_t \rangle$, the loss vectors need to be estimated from data and it is these estimators that will be used by the mirror descent algorithm. Because estimation of y_t is only possible using randomization, the algorithm cannot play the suggested action of mirror descent, but instead plays a distribution over actions with the same mean as the proposed action. Since the losses are linear, the expected additional regret by playing according to the distribution vanishes. The algorithm is summarized in Algorithm 15. Notice we have switched to capital letters because of the randomization.

```

1: Input Legendre potential  $F$  with domain  $\mathcal{D}$ , action set  $\mathcal{A}$  and learning rate
    $\eta > 0$ 
2: Choose  $\bar{A}_1 = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} F(a)$ 
3: for  $t = 1, \dots, n$  do
4:   Choose measure  $P_t$  on  $\mathcal{A}$  with mean  $\bar{A}_t$ 
5:   Sample action  $A_t$  from  $P_t$  and observe  $\langle A_t, y_t \rangle$ 
6:   Compute estimate  $\hat{Y}_t$ 
7:   Let  $\bar{A}_{t+1} = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} \eta \langle a, \hat{Y}_t \rangle + D_F(a, \bar{A}_t)$ 
8: end for
```

Algorithm 15: Online stochastic mirror descent

THEOREM 28.3 *Provided that $\mathbb{E}[\hat{Y}_t | \bar{A}_t] = y_t$ and $a \in \mathcal{A}$ and $\mathcal{A} \subseteq \operatorname{cl}(\mathcal{D})$, then*

$$R_n(a) \leq \mathbb{E} \left[\sum_{t=1}^n \langle \bar{A}_t - \bar{A}_{t+1}, \hat{Y}_t \rangle + \frac{F(a) - F(\bar{A}_1)}{\eta} - \frac{1}{\eta} \sum_{t=1}^n D(\bar{A}_{t+1}, \bar{A}_t) \right].$$

Furthermore, letting $\tilde{A}_{t+1} = \operatorname{argmin}_{a \in \mathcal{D}} \eta \langle a, \hat{Y}_t \rangle + D_F(a, \bar{A}_t)$ and assuming that $\eta \hat{Y}_t + \nabla F(x) \in \mathcal{D}^$ for all $x \in \mathcal{A}$ almost surely, then*

$$R_n \leq \frac{\operatorname{diam}_F(\mathcal{A})}{\eta} + \frac{1}{\eta} \sum_{t=1}^n \mathbb{E} [D(\bar{A}_t, \tilde{A}_{t+1})].$$

Proof Using the definition of the algorithm and the assumption that \hat{Y}_t is

unbiased given \bar{A}_t and P_t has mean \bar{A}_t leads to

$$\mathbb{E}[\langle A_t, y_t \rangle] = \mathbb{E}[\langle \bar{A}_t, y_t \rangle] = \mathbb{E}[\mathbb{E}[\langle \bar{A}_t, y_t \rangle | \bar{A}_t]] = \mathbb{E}\left[\mathbb{E}\left[\langle \bar{A}_t, \hat{Y}_t \rangle | \bar{A}_t\right]\right],$$

where the last equality used the linearity of expectations. Hence,

$$R_n(x) = \mathbb{E}\left[\sum_{t=1}^n \langle A_t, y_t \rangle - \langle x, y_t \rangle\right] = \mathbb{E}\left[\sum_{t=1}^n \langle \bar{A}_t - x, \hat{Y}_t \rangle\right],$$

which is the expected random regret of mirror descent on the recursively constructed sequence \hat{Y}_t . The result follows from Theorem 28.1 and the note at the end of the last chapter that says that this theorem continues to hold even for recursively constructed sequences. \square

The same style of proof also works for follow the regularized leader.

28.4 The unit ball

In the previous chapter we showed that continuous exponential weights on the unit ball has a regret of

$$R_n = O(d\sqrt{n \log(n)}).$$

The reader now knows that this is a version of mirror descent with the negentropy potential. Somewhat surprisingly, the dependence on the dimension can be reduced to \sqrt{d} using a more carefully chosen potential. For the remainder of this section let $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ be the standard Euclidean ball. In order to instantiate the mirror descent algorithm for bandits we need a potential, sampling rule, an unbiased estimator and a learning rate. We start with the sampling rule and estimator. Recall that in round t we need to choose a distribution on \mathcal{A} with mean \bar{A}_t and sufficient variability that the variance of the estimator is not too large. Let $E_t \in \{0, 1\}$ satisfy $\mathbb{E}_t[E_t] = (1 - \|\bar{A}_t\|)$ and U_t be an independent and uniformly distributed on $\{\pm e_1, \dots, \pm e_d\}$. Then let

$$A_t = E_t U_t + \frac{(1 - E_t)\bar{A}_t}{\|\bar{A}_t\|}.$$

Then the sampling distribution is just the law of A_t , which clearly satisfies $\mathbb{E}_t[A_t] = \bar{A}_t$. For the estimator we use a variant of the importance-weighted estimator from the last chapter:

$$\hat{Y}_t = \frac{dE_t A_t \langle A_t, y_t \rangle}{1 - \|\bar{A}_t\|}. \quad (28.10)$$

The reader can check for themselves that this estimator is unbiased. Next we inspect the contents of our magicians hat and select the potential

$$F(a) = -\log(1 - \|a\|) - \|a\|.$$

There is one more modification. Rather than instantiating mirror descent with action-set \mathcal{A} , we use $\tilde{\mathcal{A}} = \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$ where $r < 1$ is a radius to be tuned subsequently. The reason for this modification is to control the variance of the estimator in Eq. (28.10), which blows up as \bar{A}_t gets close to the boundary. Note that the exploration means that the algorithm often plays actions that are not in $\tilde{\mathcal{A}}$, but mirror descent always chooses $\bar{A}_t \in \tilde{\mathcal{A}}$.

THEOREM 28.4 *If Algorithm 15 is run using the sampling rule, estimator and potential as described above and the learning rate is $\eta = \sqrt{\log(n)/(3dn)}$ and $r = 1 - 2\eta d$. Then*

$$R_n \leq 2\sqrt{3nd\log(n)}.$$

Proof The first step is to bound the conditional variance of the estimator.

$$\mathbb{E}_t [\|\hat{Y}_t\|^2] = \frac{d^2}{(1 - \|\bar{A}_t\|)^2} \mathbb{E}_t [E_t A_t^\top A_t \langle A_t, y_t \rangle^2] = \frac{d\|y_t\|^2}{1 - \|\bar{A}_t\|} \leq \frac{d}{1 - \|\bar{A}_t\|}. \quad (28.11)$$

Turning our attention towards the properties of the potential. An easy calculation shows that

$$\nabla F(a) = \frac{a}{1 - \|a\|}, \quad F^*(u) = -\log(1 + \|u\|) + \|u\|, \quad \nabla F^*(u) = \frac{u}{1 + \|u\|}.$$

Since the domain of F is $\mathcal{D} = \{x : \|x\|_2 < 1\}$ it follows that $\nabla F(\mathcal{D}) = \mathbb{R}^d$, which means that Eq. (28.5) is satisfied. We now use the fact that $\log(x) \geq x - x^2$ for all $x \geq -1/2$, which means that if $(\|u\| - \|v\|)/(1 + \|v\|) \geq -1/2$, then

$$\begin{aligned} D_{F^*}(u, v) &= -\log\left(\frac{1 + \|u\|}{1 + \|v\|}\right) + \|u\| - \|v\| - \frac{1}{1 + \|v\|} \langle v, u - v \rangle \\ &= \frac{1}{1 + \|v\|} \left(\|u\| - \|v\| + \|v\|\|u\| - \langle v, u \rangle - (1 + \|v\|) \log\left(1 + \frac{\|u\| - \|v\|}{1 + \|v\|}\right) \right) \\ &\leq \frac{1}{1 + \|v\|} \left(\|v\|\|u\| - \langle v, u \rangle + \frac{(\|u\| - \|v\|)^2}{1 + \|v\|} \right) \\ &\leq \frac{1}{1 + \|v\|} (\|v\|\|u\| - \langle v, u \rangle + \|u\|^2 + \|v\|^2 - \|u\|\|v\|) \\ &\leq \frac{1}{1 + \|v\|} (\|u\|^2 + \|v\|^2 - 2\langle v, u \rangle) = \frac{\|v - u\|^2}{1 + \|v\|}. \end{aligned}$$

Let \tilde{A}_{t+1} be defined as in the statement of Theorem 28.3. By the triangle inequality we have

$$\frac{\|\nabla F(\tilde{A}_{t+1})\| - \|\nabla F(\bar{A}_t)\|}{1 + \|\nabla F(\bar{A}_t)\|} \geq -\frac{\|\nabla F(\tilde{A}_{t+1}) - \nabla F(\bar{A}_t)\|}{1 + \|\nabla F(\bar{A}_t)\|} \geq -\eta \|\hat{Y}_t\| \geq -\frac{1}{2},$$

where the last inequality follows since $\eta \|\hat{Y}_t\| \leq \eta d/(1 - r) \leq 1/2$. By the remarks

at the end of the Section 28.2 and Eq. (28.11) leads to

$$\begin{aligned}\mathbb{E} [D_F(\bar{A}_t, \tilde{A}_{t+1})] &= \mathbb{E} [D_{F^*}(\nabla F(\tilde{A}_{t+1}), \nabla F(\bar{A}_t))] \\ &\leq \mathbb{E} \left[\frac{\|\nabla F(\tilde{A}_{t+1}) - \nabla F(\bar{A}_t)\|^2}{1 + \|\nabla F(\bar{A}_t)\|} \right] \\ &= \mathbb{E} \left[\eta^2 (1 - \|\bar{A}_t\|) \|\hat{Y}_t\|^2 \right] \leq \eta^2 d.\end{aligned}$$

By Theorem 28.3 for any $a \in \tilde{\mathcal{A}}$ and using the fact that $\bar{A}_1 = 0$ and $F(\bar{A}_1) = 0$,

$$R_n(a) \leq \frac{F(a) - F(\bar{A}_1)}{\eta} + \eta nd \leq \frac{1}{\eta} \log \left(\frac{1}{1 - \|a\|} \right) + \eta nd.$$

Let $a^* \in \operatorname{argmin}_{a \in \mathcal{A}} \sum_{t=1}^n \langle a, y_t \rangle$, then

$$\begin{aligned}R_n(a) &= \mathbb{E} \left[\sum_{t=1}^n \langle A_t - a^*, y_t \rangle \right] = \mathbb{E} \left[\sum_{t=1}^n \langle A_t - ra^*, y_t \rangle \right] + \sum_{t=1}^n \langle ra^* - a^*, y_t \rangle \\ &\leq \frac{1}{\eta} \log \left(\frac{1}{1 - \|a\|} \right) + \eta nd + n(1 - r) \leq \frac{1}{\eta} \log(n) + 3\eta nd,\end{aligned}$$

which completes the proof. \square

Surprisingly this is smaller than the regret that we got for stochastic bandits by a factor of at least \sqrt{d} . There is no contradiction because the adversarial and stochastic linear bandit models are actually quite different, a topic to which the next chapter is dedicated.

28.5 Notes

- 1 Finding a_{t+1} for both mirror descent and follow the regularized leader requires solving a convex optimization problem. Provided the dimension is not too large and the action-set and potential are reasonably nice, there exist practical approximation algorithms for this problem. The two-step process described in Eqs. (28.6) and (28.7) is sometimes an easier way to go. Usually (28.6) can be solved analytically while (28.7) can be quite expensive. In some important special cases, however, the projection step can be written in closed form or efficiently approximated.
- 2 We saw that mirror descent with a carefully chosen potential function achieves $O(\sqrt{dn \log(n)})$ regret on the ℓ_2 -ball. On the ℓ_∞ ball (hypercube) the optimal regret is $O(d\sqrt{n})$. Interestingly, as n tends to infinity the optimal dependence on the dimension is either d or \sqrt{d} with a complete classification given by Bubeck et al. [2018].
- 3 Adversarial linear bandits on a simplex are equivalent to finite-armed adversarial bandits with d arms. Yet another well-chosen potential function leads to an algorithm with regret $R_n = O(\sqrt{dn})$, which matches the lower bound and

shaves a factor of $\sqrt{\log d}$ from the upper bounds presented in Chapters 11 and 12. For more details see Exercise 28.10.

- 4 Both mirror descent and follow the regularized leader depend on a carefully chosen potential function. Currently there is no characterization of exactly what this potential should be or how to find it. At least in the full information setting there are quite general universality results showing that if a certain regret is achievable by some algorithm, then that same regret is nearly achievable by mirror descent with some potential [Srebro et al., 2011]. In practice this result is not useful for constructing new potential functions, however. There have been some attempts to develop ‘universal’ potential functions that exhibit nice behavior for any action sets [Bubeck et al., 2015b, and others]. These can be useful, but as yet we do not know precisely what properties are crucial, especially in the bandit case.
- 5 When the horizon is unknown the learning rate cannot be tuned ahead of time. One option is to apply the doubling trick. A more elegant solution is to use a decreasing schedule of learning rates. This requires an adaptation of the proofs of Theorems 28.1 and 28.2, which we outline in Exercises 28.7 and 28.8. This is one situation where mirror descent and follow the regularized leader are not the same and where results slightly favour the latter algorithm.
- 6 In much of the literature the potential is chosen in such a way that mirror descent and follow the regularized leader are the same algorithm. For historical reasons the name mirror descent is more commonly used in the bandit community. We encourage the reader to keep both algorithms in mind, since the analysis of one-or-other can sometimes be slightly easier. Note that **lazy mirror descent** is a variant of mirror descent that is equivalent to follow the regularized leader for all Legendre potentials [Hazan, 2016].
- 7 We mentioned that for online linear optimization the mirror descent algorithm also works when y_1, \dots, y_n are chosen nonobliviously. This does not translate to the bandit setting for a subtle reason. Let $\hat{R}_n(a) = \sum_{t=1}^n \langle A_t - a, y_t \rangle$ be the random regret so that

$$R_n = \mathbb{E} \left[\max_{a \in \mathcal{A}} \hat{R}_n(a) \right] = \mathbb{E} \left[\sum_{t=1}^n \langle A_t, y_t \rangle - \max_{a \in \mathcal{A}} \sum_{t=1}^n \langle a, y_t \rangle \right].$$

The second sum is constant when the losses are oblivious, which means the maximum can be brought outside the expectation, which is not true if the loss vectors are nonoblivious. It is still possible to bound the expected loss relative to a fixed comparator a so that

$$R_n(a) = \mathbb{E} \left[\sum_{t=1}^n \langle A_t - a, y_t \rangle \right] \leq B,$$

where B is whatever bound obtained from the analysis presented above. A

little rewriting shows that

$$R_n = \mathbb{E} \left[\max_{a \in \mathcal{A}} \hat{R}_n(a) \right] = B + \mathbb{E} \left[\max_{a \in \mathcal{A}} R_n(a) \right] - \max_{a \in \mathcal{A}} \mathbb{E} [R_n(a)] .$$

The difference in expectations can be bounded using tools from empirical process theory, but the resulting bound is only $O(\sqrt{n})$ if $\mathbb{V}[\hat{R}_n(a)] = O(n)$. In general, however, the variance can be as large as $n^{3/2}$ so this condition must be checked for each proposed policy. We emphasize again that the nonoblivious regret is a strange measure because it does not capture the reactive nature of the environment. The details of the application of empirical process theory is beyond the scope of this book. For an introduction to that topic we recommend the books by [Vaart and Wellner \[1996\]](#), [Dudley \[2014\]](#) and [van de Geer \[2000\]](#).

- 8 The price of bandit information on the unit ball is an extra $\sqrt{d \log(n)}$ (compare Proposition 28.1 and Theorem 28.4). Except for log factors this is also true for the simplex (Proposition 28.2 and Note 3). One might wonder if the difference is always about \sqrt{d} , but this is not true in general. The price of bandit information can be as high as $\Theta(d)$. Overall the dimension dependence in the regret in terms of the action set is still not well understood except for special cases.
- 9 The poor behavior of follow the leader in the full information setting depends on (a) the environment being adversarial rather than stochastic and (b) the action set having sharp corners. When either of these factors is missing, follow the leader is a reasonable choice [[Huang et al., 2017b](#)]. Note that with bandit feedback the failure is primarily due to a lack of exploration (Exercises 4.5 and 4.6).
- 10 A simultaneous-action zero sum game is a game between two players. As the name suggests, both players simultaneously choose a distribution $p \in \mathcal{P}_{K-1}$ and $q \in \mathcal{P}_{E-1}$ respectively. The loss to the first player is $\langle p, Gq \rangle$ where $G \in [0, 1]^{K \times E}$. The loss for the second player is $-\langle p, Gq \rangle$, which means that the sum of the players losses is zero ('zero sum'). The minimax theorem states that

$$\min_p \max_q \langle p, Gq \rangle = \max_q \min_p \langle p, Gq \rangle$$

Let p^* be the minimizing distribution on the left-hand side and q^* be the maximizing distribution on the right-hand side. Then the pair (p^*, q^*) is called a **Nash equilibrium** of the game, which satisfies the condition that either player can announce their strategy in advance without consequences.

28.6 Bibliographic remarks

The results in this chapter come from a wide variety of sources. The online convex optimization framework is due to [Zinkevich \[2003\]](#), where online gradient descent is introduced and analyzed. Mirror descent was first developed by [Nemirovski \[1979\]](#) and [Nemirovski and Yudin \[1983\]](#) for classical optimization while 'follow

the regularized leader' appears first in the work by Shalev-Shwartz [2007], Shalev-Shwartz and Singer [2007]. An implicit form of regularization is to add a perturbation of the losses, leading to the 'follow the perturbed leader' algorithm [Hannan, 1957, Kalai and Vempala, 2005a], which is further explored in the context of combinatorial bandit problems in Chapter 30 (and see also Exercise 11.7). Readers interested in an overview of online learning will like the short book by Hazan [2016] while the book by Cesa-Bianchi and Lugosi [2006] has a little more depth (but is also ten years older). As far as we know, the first application of mirror descent to bandits was by Abernethy et al. [2008]. Since then the idea has been used extensively with some examples by Audibert et al. [2013], Abernethy et al. [2015], Bubeck et al. [2018]. Mirror descent has been adapted in a generic way to prove high probability bounds by Abernethy and Rakhlin [2009]. The reader can find (slightly) different proofs of some mirror descent results in the book by Bubeck and Cesa-Bianchi [2012]. The result for the unit ball are from a paper by Bubeck et al. [2012]. Mirror descent can be generalized to Banach spaces. For details see the article by Sridharan and Tewari [2010].

28.7 Exercises

28.1 Show that if F is Legendre with domain $\mathcal{D} \subseteq \mathcal{A} \subset \mathbb{R}^d$ then the minimizer of $\Phi_t(a) = \eta \langle a, y_t \rangle + D_F(a, a_t) = \eta \langle a, y_t \rangle + F(a) - F(a_t) - \langle \nabla F(a_t), a - a_t \rangle$ over \mathcal{A} belongs to the interior of \mathcal{D} and at the minimizer a_{t+1} , $\nabla \Phi_t(a_{t+1}) = 0$ holds.

28.2 Let $\mathcal{A} = [-1, 1]$ and let $y_1 = 1/2$ and $y_s = 1$ for odd $s > 1$ and $y_s = -1$ for even $s > 1$.

- (a) Recall that follow the leader (without regularization) chooses $a_t = \operatorname{argmin}_a \sum_{s=1}^{t-1} \langle a, y_s \rangle$. Show that this algorithm suffers linear regret.
- (b) Implement follow the regularized leader or mirror descent on this problem with quadratic potential $F(a) = a^2$ and plot a_t as a function of time.

28.3 Prove Corollary 28.1.

28.4 Let $\mathcal{A} = \mathcal{P}_{K-1}$ be the simplex, F the unnormalized negentropy potential and $\eta > 0$. Let $P_1 = \operatorname{argmin}_{p \in \mathcal{A}} F(p)$ and for $t \geq 1$ let

$$P_{t+1} = \operatorname{argmin}_{p \in \mathcal{A}} \eta \langle p, \hat{Y}_t \rangle + D_F(p, P_t),$$

where $\hat{Y}_{ti} = \mathbb{I}\{A_t = i\} y_{ti}/P_{ti}$ and A_t is sampled from P_t .

- (a) Show that the resulting algorithm is exactly Exp3 from Chapter 11.
- (b) What happens if you replace mirror descent by follow the regularized leader? That is, if $P_t = \operatorname{argmin}_{p \in \mathcal{A}} \sum_{s=1}^{t-1} \langle p, \hat{Y}_s \rangle + F(p)$.

28.5 Continuing on from the last exercise, in this exercise you will show that

the tools in this chapter not only lead to the same algorithm, but also the same bounds.

- (a) Let $\tilde{P}_{t+1} = \operatorname{argmin}_{p \in [0, \infty)^K} \eta \langle p, \hat{Y}_t \rangle + D_F(p, P_t)$. Show both relations in the following display:

$$D_F(P_t, \tilde{P}_{t+1}) = \sum_{i=1}^K P_{ti} \left(\exp(-\eta \hat{Y}_{ti}) - 1 + \eta \hat{Y}_{ti} \right) \leq \frac{\eta^2}{2} \sum_{i=1}^K P_{ti} \hat{Y}_{ti}^2.$$

- (b) Show that $\frac{1}{\eta} \mathbb{E} \left[\sum_{t=1}^n D_F(P_t, \tilde{P}_{t+1}) \right] \leq \frac{\eta n K}{2}$.

- (c) Show that $\operatorname{diam}_F(\mathcal{P}_{K-1}) = \log(K)$.

- (d) Conclude that for appropriately tuned $\eta > 0$ the regret of Exp3 satisfies,

$$R_n \leq \sqrt{2nK \log(K)}.$$

28.6 Prove Theorem 28.2.

28.7 Let \mathcal{A} be closed and convex and $y_1, \dots, y_n \in \mathcal{L} \subseteq \mathbb{R}^d$. Let F be Legendre with domain \mathcal{D} and assume that $\mathcal{A} \subseteq \operatorname{cl}(\mathcal{D})$ and that Eq. (28.5) holds. Let $\eta_0, \eta_1, \dots, \eta_n$ be a sequence of learning rates where we assume that $\eta_0 = \infty$ and $a_1 = \operatorname{argmin}_{a \in \mathcal{A}} F(a)$ and

$$\begin{aligned} \tilde{a}_{t+1} &= \operatorname{argmin}_{a \in \mathcal{D}} \eta_t \langle a, y_t \rangle + D_F(a, a_t), \\ a_{t+1} &= \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} D_F(a, \tilde{a}_{t+1}). \end{aligned}$$

Show that for all $a \in \mathcal{A}$:

- (a) $R_n(a) = \sum_{t=1}^n \langle a_t - a, y_t \rangle \leq \sum_{t=1}^n \frac{D_F(a_t, \tilde{a}_{t+1})}{\eta_t} + \sum_{t=1}^n \frac{D_F(a, a_t) - D_F(a, \tilde{a}_{t+1})}{\eta_t}$.
- (b) $R_n(a) \leq \sum_{t=1}^n \frac{D_F(a_t, \tilde{a}_{t+1})}{\eta_t} + \sum_{t=1}^n D_F(a, a_t) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right)$.

28.8 Like in the previous exercise let \mathcal{A} be closed and convex and $y_1, \dots, y_n \in \mathcal{L} \subseteq \mathbb{R}^d$. Let F_1, \dots, F_n be a sequence of Legendre functions where $\operatorname{dom}(F_t) = \mathcal{D}_t$ and $\mathcal{A} \subseteq \operatorname{cl}(\mathcal{D}_t)$ for all t . Let $\Phi_t(a) = F_t(a) + \sum_{s=1}^t \langle a, y_s \rangle$ and $a_t = \operatorname{argmin}_{a \in \mathcal{A}} \Phi_{t-1}(a)$. Show that

$$\begin{aligned} R_n(a) &\leq \sum_{t=1}^n (\langle a_t - a_{t+1}, y_t \rangle - D_{F_{t-1}}(a_{t+1}, a_t)) \\ &\quad + F_n(a) - F_0(a_1) + \sum_{t=1}^n (F_{t-1}(a_{t+1}) - F_t(a_{t+1})). \end{aligned}$$

28.9 Consider the finite-armed adversarial bandit problem described in Chapter 11 where the adversary chooses y_1, \dots, y_n with $y_t \in [0, 1]^K$. Let

$P_t \in \mathcal{P}_{K-1}$ be defined by

$$P_{ti} = \frac{\exp\left(-\eta_{t-1} \sum_{s=1}^{t-1} \hat{Y}_{si}\right)}{\sum_{j=1}^K \exp\left(-\eta_{t-1} \sum_{s=1}^{t-1} \hat{Y}_{sj}\right)},$$

where η_0, η_1, \dots is an infinite sequence of learning rates and $\hat{Y}_{ti} = \mathbb{I}\{A_t = i\} y_{ti}/P_{ti}$ and A_t is sampled from P_t .

- (a) Let $\mathcal{A} = \mathcal{P}_{K-1}$ be the simplex, F be the negentropy potential, $F_t(p) = F(p)/\eta_t$ and $\Phi_t(p) = F(p)/\eta_t + \sum_{s=1}^t \langle p, \hat{Y}_s \rangle$. Show that P_t is the choice of follow the regularized leader with potentials (F_t) and losses (\hat{Y}_t).
- (b) Let $P \in \mathcal{P}_{K-1}$ be the standard basis vector with $P_i = 1$ for $i = \operatorname{argmin}_j \sum_{t=1}^n y_{tj}$. Use the fact that \hat{Y}_t is an unbiased estimate of y_t and Exercise 28.8 to show that

$$\begin{aligned} R_n &\leq \mathbb{E} \left[\sum_{t=1}^n \langle P_t - P_{t+1}, \hat{Y}_t \rangle - D_{F_t}(P_t, P_{t+1}) \right] \\ &\quad + F_n(P) - F_0(P_1) + \sum_{t=1}^n (F_{t-1}(P_{t+1}) - F_t(P_{t+1})). \end{aligned}$$

- (c) Assume that (η_t) is decreasing and show that

$$F_n(P) - F_0(P_1) + \sum_{t=1}^n (F_{t-1}(a_{t+1}) - F_t(a_{t+1})) \leq \frac{\log(K)}{\eta_n}.$$

- (d) Use Theorem 26.5 in combination with the facts that $\hat{Y}_{ti} \geq 0$ for all i and $\hat{Y}_{ti} = 0$ unless $A_t = i$ to show that

$$\langle P_t - P_{t+1}, \hat{Y}_t \rangle - D_{F_t}(P_t, P_{t+1}) \leq \frac{\eta_t}{2P_{tA_t}}.$$

- (e) Prove that $R_n \leq \frac{\log(K)}{\eta_n} + \frac{K}{2} \sum_{t=1}^n \eta_t$.

- (f) Choose $\eta_0, \eta_1, \eta_2, \dots$ so that $R_n \leq 2\sqrt{nK \log(K)}$.

28.10 Let $\mathcal{A} = \mathcal{P}_{K-1}$ be the simplex and assume $y_t \in \mathcal{A}^\circ$ for all t and let $F(a) = -2 \sum_{i=1}^d \sqrt{a_i}$.

- (a) Show that $F^*(u) = -\sum_{i=1}^K u_i^{-1}$ whenever $u \in (-\infty, 0]^K$.
- (b) Show that for $u, v \in (-\infty, 0]^K$,

$$D_{F^*}(u, v) = -\sum_{i=1}^K \frac{(u_i - v_i)^2}{u_i v_i^2}.$$

- (c) Show that $\operatorname{diam}_F(\mathcal{A}) \leq 2\sqrt{d}$.

- (d) Let A_t be chosen so that $\mathbb{P}(A_t = i | \bar{A}_t) = \bar{A}_{ti}$ and \hat{Y}_t be the importance-weighted estimator

$$\hat{Y}_{ti} = \frac{\mathbb{I}\{A_t = i\} y_{ti}}{\bar{A}_{ti}}.$$

- (e) Show that $\tilde{A}_{t+1,A_t} \leq \bar{A}_{t+1,A_t}$ and $\tilde{A}_{t+1,i} = \bar{A}_{t+1,i}$ for $i \neq A_t$.
(f) Show that $\mathbb{E}[D_{F^*}(\nabla F(\tilde{A}_{t+1}), \nabla F(\bar{A}_t))] \leq \eta^2 \sqrt{d}$.
(g) Conclude that the regret of mirror descent with this potential is bounded by

$$R_n \leq \sqrt{8dn}.$$

- (h) Devise an efficient implementation of this algorithm.



The algorithm in the above exercise is called the **implicitly normalized forecaster** and was introduced by [Audibert and Bubeck \[2009\]](#). At first it went unnoticed that this algorithm was an instance of mirror descent and the proof was consequentially much more complicated. More details are in the book by [Bubeck and Cesa-Bianchi \[2012\]](#).

28.11 Let F be the unnormalized negentropy potential and consider online mirror descent with $\mathcal{A} = \mathcal{P}_{d-1}$ and loss vectors y_1, \dots, y_n chosen from the hypercube: $y_t \in [0, 1]^d$. Prove that $R_n \leq \sqrt{2n \log(K)}$.

28.12 Prove the minimax theorem described in Note 10.



Let $p_1, \dots, p_n \in \mathcal{P}_{K-1}$ be the choices of mirror descent with unnormalized negentropy potential when the losses y_1, \dots, y_n are given by $y_t = Gq_t$ and $q_t = \operatorname{argmax}_q \langle p_t, Gq \rangle$. Then use the result from Exercise 28.11. The minimax theorem is due to [von Neumann \[1928\]](#).

29 The Relation Between Adversarial and Stochastic Linear Bandits

As we have seen in the preceding chapters, adversarial and stochastic linear bandits do share certain similarities. For example, the least squares estimator plays a fundamental role in both, as does the machinery from experimental design for optimizing the exploration distribution. There are also surprising differences, however. Theorem 24.2 shows that the regret for stochastic linear bandits on a ball is lower bounded by $\Omega(d\sqrt{n})$, while for the adversarial bandits the upper bound is $O(\sqrt{dn})$ as shown in Theorem 28.4. As we will keep referring to these results, we added Table 29.1 to summarize the situation. We hope the reader is at least mildly surprised that the regret upper for the adversarial environment is of lower order than the regret lower bound for the stochastic environment. After all, was not the purpose of working with adversarial environments to enlarge the scope of algorithms beyond stochastic environments? The purpose of this chapter is to explain why our intuition fails us in this case.

To make the notation consistent we present the stochastic and adversarial linear bandit frameworks again, but this time using losses for both. Let $\mathcal{A} \subset \mathbb{R}^d$ be the action set. In each round the learner should choose $A_t \in \mathcal{A}$ and receives the loss Y_t , where

$$Y_t = \langle A_t, \theta \rangle + \eta_t, \quad (\text{Stochastic setting}) \quad (29.1)$$

$$Y_t = \langle A_t, \theta_t \rangle, \quad (\text{Adversarial setting}) \quad (29.2)$$

where $(\eta_t)_t$ is a sequence of independent and identically distributed 1-subgaussian random variables and $(\theta_t)_t$ is a sequence of loss vectors chosen by the adversary. As noted earlier, the assumptions on the noise can be relaxed significantly. For example, if $\mathcal{F}_t = \sigma(A_1, Y_1, \dots, A_t, Y_t, A_{t+1})$, then the results of the previous chapters hold as soon as $\eta_t | \mathcal{F}_{t-1} \sim \text{subG}(1)$. The expected regret for the two

	Stochastic environment	Adversarial environment
Regret	$\Omega(d\sqrt{n})$	$O(\sqrt{dn})$

Table 29.1 The behavior of regret as a function of dimension d and number of rounds n for linear bandits when the action set $\mathcal{A} = B_2^d$ is the d -dimension Euclidean ball.

cases are defined as follows:

$$\begin{aligned} R_n &= \sum_{t=1}^n \mathbb{E} [\langle A_t, \theta \rangle] - n \inf_{a \in \mathcal{A}} \langle a, \theta \rangle, && \text{(Stochastic setting)} \\ R_n &= \sum_{t=1}^n \mathbb{E} [\langle A_t, \theta_t \rangle] - n \inf_{a \in \mathcal{A}} \langle a, \bar{\theta}_n \rangle. && \text{(Adversarial setting)} \end{aligned}$$

In the last display, $\bar{\theta}_n = \frac{1}{n} \sum_{t=1}^n \theta_t$ is the average of the loss vectors chosen by the adversary. We chose to write the adversarial form with the help of the average loss vector to emphasize the similarity between the two settings.

29.1 Reducing stochastic linear bandits to adversarial linear bandits

To formalize the intuition that adversarial environments are harder than stochastic environments one may try to find a **reduction** where learning in the stochastic environments is reduced to learning in the adversarial algorithms. Here, reducing problem E ('easy') to problem H ('hard') just means that we can use algorithms designed for problem H to solve instances of problem E. In order to do this we need to transform instances of problem E into instances of problem H and translate back the actions of algorithms to actions for problem E. To get a regret bound for problem E from regret bound for problem H, one needs to ensure that the losses translate properly between the problem classes.

Of course, based on our previous discussion we know that if there is a reduction from stochastic linear bandits to adversarial linear bandits then somehow the adversarial problem must change so that no contradiction is created in the curious case of the unit ball. To be able to use an adversarial algorithm in the stochastic environment, we need to specify a sequence $(\theta_t)_t$ so that the adversarial feedback matches the stochastic one. Comparing Eq. (29.1) and Eq. (29.2), we can see that the crux of the problem is incorporating the noise η_t into θ_t while satisfying the other requirements. One simple way of doing this is by introducing an extra dimension for the adversarial problem.

In particular, suppose that the stochastic problem is d -dimensional so that $\mathcal{A} \subset \mathbb{R}^d$. For the sake of simplicity, assume furthermore that the noise and parameter vector satisfy $|\langle \mathcal{A}, \theta \rangle + \eta_t| \leq 1$ almost surely and that $a_* = \operatorname{argmin}_{a \in \mathcal{A}} \langle a, \theta \rangle$ exists. Then define $\mathcal{A}_{\text{aug}} = \{(a, 1) : a \in \mathcal{A}\} \subset \mathbb{R}^{d+1}$ and let the adversary choose $\theta_t = (\theta, \eta_t) \in \mathbb{R}^{d+1}$. The reduction is now straightforward:

- 1 Initialize adversarial bandit policy with action set \mathcal{A}_{aug} .
- 2 Collect action $A'_t = (A_t, 1)$ from the policy.
- 3 Play A_t and observe loss Y_t .
- 4 Feed Y_t to the adversarial bandit policy and repeat from step 2.

Suppose the adversarial policy guarantees a bound B_n on the expected regret:

$$R'_n = \mathbb{E} \left[\sum_{t=1}^n \langle A'_t, \theta_t \rangle - \inf_{a' \in \mathcal{A}_{\text{aug}}} \sum_{t=1}^n \langle a', \theta_t \rangle \right] \leq B_n.$$

Let $a'_* = (a_*, 1)$. Note that for any $a' = (a, 1) \in \mathcal{A}_{\text{aug}}$, $\langle A_t, \theta \rangle - \langle a, \theta \rangle = \langle A'_t, \theta_t \rangle - \langle a', \theta_t \rangle$ and thus adversarial regret, and eventually B_n , will upper bound the stochastic regret:

$$\mathbb{E} \left[\sum_{t=1}^n \langle A_t, \theta \rangle \right] - n \langle a_*, \theta \rangle = \mathbb{E} \left[\sum_{t=1}^n \langle A'_t, \theta_t \rangle - n \langle a'_*, \bar{\theta}_n \rangle \right] \leq R'_n \leq B_n.$$

Therefore the expected regret in the stochastic bandit is also at most B_n . We have to emphasize that this reduction changes the geometry of the decision sets for both the learner and the adversary. For example, if $\mathcal{A} = B_2^d$ is the unit ball, then neither \mathcal{A}_{aug} nor its polar $\mathcal{A}_{\text{aug}}^\circ$ are unit balls. It does not seem like this should make much difference, but at least in the case of the ball, from our $\Omega(d\sqrt{n})$ lower bound on the regret for the stochastic case, we see that the changed geometry must make the adversary more powerful. This reinforces the importance of the geometry of the action set, which we have already seen in the previous chapter.

While the reduction shows one way to use adversarial algorithms in stochastic environments, the story seems to be unfinished. When facing a linear bandit problem with some action set \mathcal{A} , the user is forced to make a choice of whether to believe the environment to be stochastic. Strangely enough, if the environment is believed to be stochastic, the recommendation seems to be to run one's favorite adversarial linear bandit algorithm on the *augmented* action set. What if the environment may or may not be stochastic? One can still try to run the adversarial linear bandit algorithm with no changes. At present we cannot guarantee that this lead to a small regret. In fact, the regret may get larger than what it needs to be. For example, if the mirror descent algorithm of the last chapter is run on a stochastic environment with the learning rate of the algorithm set as recommended in Theorem 28.4, the regret upper bound increases to $\tilde{O}(d^2\sqrt{n})$. We believe that this increase of the regret is real. By tuning the learning rate, the regret can be brought back to $\tilde{O}(d\sqrt{n})$.



We see a case here when the cost of using an algorithm prepared to deal with a larger class of environments pays a nontrivial cost for its increased robustness. At least as far as minimax regret is concerned, this was not the case for finite-armed bandits.

The real reason for all these discrepancies is that that the adversarial linear bandit model is better viewed as relaxation of another class of stochastic linear bandits, which we discuss in the next section.

29.2 Stochastic linear bandits with parameter noise

Another way to relate the adversarial and stochastic linear bandit frameworks is to start from the adversarial model and add stochasticity by assuming that θ_t is chosen from some fixed distribution $\nu \in \mathbb{R}^d$. We call the resulting stochastic linear bandit model the **stochastic linear bandit with parameter noise**. This new problem can be trivially reduced to adversarial bandits (assuming the support of ν is bounded). In particular, there is no need to change the action sets. We note in passing that constructing a stochastic environment like this is often the way lower bounds are constructed for adversarial models.

Parameter noise environments form a subset of all possible stochastic environments. To see this, let $\theta = \int x\nu(dx)$ be the mean parameter vector under ν . Then, the loss (or reward) in round t is

$$\langle A_t, \theta_t \rangle = \langle A_t, \theta \rangle + \langle A_t, \theta_t - \theta \rangle.$$

Let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_{t-1}]$. By our assumption that ν has mean θ the second term vanishes in expectation, $\mathbb{E}_t[\langle A_t, \theta_t - \theta \rangle] = 0$. This implies that we can make a connection to the ‘vanilla’ stochastic setting by letting $\tilde{\eta}_t = \langle A_t, \theta_t - \theta \rangle$. Now consider the conditional variance of $\tilde{\eta}_t$:

$$\mathbb{V}_t[\tilde{\eta}_t] = \mathbb{E}_t[\langle A_t, \theta_t - \theta \rangle^2] = A_t^\top \mathbb{E}_t[(\theta_t - \theta)(\theta_t - \theta)^\top] A_t = A_t^\top \Sigma A_t, \quad (29.3)$$

where Σ is the covariance matrix of multivariate distribution ν . Eq. (29.3) implies that the variance of the noise $\tilde{\eta}_t$ now depends on the choice of action and in particular the noise variance scales with the length of A_t . This can make parameter noise problems easier. For example, if ν is a Gaussian with identity covariance, then $\mathbb{V}_t[\tilde{\eta}_t] = \|A_t\|_2^2$ so that long actions have more noise than short actions. By contrast, in the usual stochastic linear bandit, the variance of the noise is unrelated to the length of the action. In particular, even the noise accompanying short actions can be large. This makes quite a bit of difference in cases when the action set has both short and long actions. In the standard stochastic model, shorter actions have the disadvantage of having a worse signal-to-noise ratio, which an adversary can exploit.

This calculation also provides the reason for the different guarantees for the unit ball. For stochastic linear bandits with 1-subgaussian noise the regret is $\tilde{O}(d\sqrt{n})$ while in the last chapter we showed that for adversarial linear bandits the regret is $\tilde{O}(\sqrt{dn})$. This discrepancy is explained by the variance of the noise. Suppose that ν is supported on the unit sphere, then the eigenvalues of its covariance matrix sum to 1 and if the learner chooses A_t from the uniform probability measure μ on the sphere, then

$$\mathbb{E}[\mathbb{V}_t[\tilde{\eta}_t]] = \int a^\top \Sigma a d\mu(a) = 1/d,$$

By contrast, in the standard stochastic model with 1-subgaussian noise the predictable variation of the noise is just 1. If the adversary were allowed to choose

its loss vectors from the sphere of radius \sqrt{d} , then the expected predictable variation would be 1, matching the standard stochastic case, and the regret would scale linearly in d , which also matches the vanilla stochastic case. This example further emphasizes the importance of the assumptions that restrict the choices of the adversary.



The main takeaway of this chapter that the best way to think about the standard adversarial linear model is that it generalizes the stochastic linear bandit model under parameter noise, which is a special case stochastic linear bandits, which oftentimes is easier than the full stochastic linear bandit problem because parameter noise limits the adversary's control of the signal-to-noise ratio experienced by the learner.

29.3 Notes

- One obvious issue with the stochastic linear bandit model is that the feedbacks may not follow it! It is tempting to try and use adversarial bandits to resolve the resulting unrealizable linear bandit problem where

$$X_t = \langle A_t, \theta \rangle + \eta_t + \varepsilon(A_t),$$

with $\varepsilon : \mathcal{A} \rightarrow \mathbb{R}$ some function with small supremum norm. Because $\varepsilon(A_t)$ depends on the chosen action it is not possible to write $X_t = \langle A_t, \theta_t \rangle$ for some θ_t that does not depend on A_t . However, at least in some cases, the idea of using an adversarial linear bandit can be shown to work (cf. Exercise 29.4).

- For the reduction in Section 29.1 we assumed that $|\langle A_t, \theta \rangle + \eta_t| \leq 1$ almost surely. This is not true for many classical noise models like the Gaussian. One way to overcome this annoyance is to apply the adversarial analysis on the event that $|\langle A_t, \theta \rangle + \eta_t| \leq C$ for some constant $C > 0$ that is sufficiently large that the probability that this event occurs is high. For example, if η_t is a standard Gaussian and $\sup_{a \in \mathcal{A}} |\langle a, \theta \rangle| \leq 1$, then C may be chosen to be $1 + \sqrt{4 \log(n)}$ and the failure event that there exists a t such that $|\langle A_t, \theta \rangle + \eta_t| \geq C$ has probability at most $1/n$ by Theorem 5.1 and a union bound.
- The mirror descent analysis of adversarial linear bandits also works for stochastic bandits. Recall that mirror descent samples A_t from a distribution with a conditional mean of \bar{A}_t and suppose that $\hat{\theta}_t$ is a conditionally unbiased estimator of θ . Then the regret for a stochastic linear bandit with optimal action a^* can be rewritten as

$$R_n = \mathbb{E} \left[\sum_{t=1}^n \langle a^* - A_t, \theta \rangle \right] = \mathbb{E} \left[\sum_{t=1}^n \langle a^* - \bar{A}_t, \theta \rangle \right] = \mathbb{E} \left[\sum_{t=1}^n \langle a^* - \bar{A}_t, \hat{\theta}_t \rangle \right].$$

Except that we have switched from losses to gains, this is now in the standard format necessary for the analysis of mirror descent. In general for the stochastic

setting the covariance of the least squares estimator $\hat{\theta}_t$ will not be the same as in the adversarial setting, however, which leads to different results. When $\hat{\theta}_t$ is biased, the bias term can be incorporated into the above formula and then bounded separately.

- 4 Consider a stochastic bandit with \mathcal{A} the unit ball and $X_t = \langle A_t, \theta \rangle + \eta_t$ where $\eta_t \in [-1, 1]$ almost surely and θ is also in the unit ball. Adapting the analysis of the algorithm in Section 28.4 leads to a bound of $O(d\sqrt{n \log(n)})$. Essentially the only change is the variance calculation in Eq. (28.11), which increases by roughly a factor of d . The details of this calculation are left to you in Exercise 29.5.

29.4 Bibliographic remarks

Linear bandits on the sphere with parameter noise have been studied by [Carpentier and Munos \[2012\]](#). However they consider the case where the action-set is the sphere and the components of the noise are independent so that $X_t = \langle A_t, \theta + \eta_t \rangle$ where the coordinates of $\eta_t \in \mathbb{R}^d$ are independent with unit variance. In this case the predictable variation is $\mathbb{V}[X_t | A_t] = \sum_{i=1}^d A_{ti}^2 = 1$ for all actions A_t and the parameter noise is equivalent to the standard model. We are not aware of any systematic studies of parameter noise in the stochastic setting. With only a few exceptions, the impact on the regret of the action-set and adversaries choices is not well understood beyond the case where \mathcal{A} is an ℓ_p -ball and the adversary chooses losses from the polar of \mathcal{A} . A variety of lower bounds illustrating the complications are given by [Shamir \[2015\]](#). Perhaps the most informative is the observation that obtaining $O(\sqrt{dn})$ regret is not possible when $\mathcal{A} = \{a + x : \|x\|_2 \leq 1\}$ is a shifted unit ball with $a = (2, 0, \dots, 0)$, which also follows from our reduction in Section 29.1.

29.5 Exercises

29.1 Complete the claims made at the end of Section 29.1. In particular, show that the bandit algorithm of Theorem 28.4 achieves an $O(d^2\sqrt{n})$ expected regret when applied to a stochastic bandit problem where the noise $(\eta_t)_t$ sequence is bounded by a constant and when used with the learning rate as described in that theorem. Further, show that by an appropriate adjustment of the learning rate, the regret can be improved to $O(d\sqrt{n})$.

29.2 Let $\mathcal{A} \subset \mathcal{R}^d$ be an action set. Take an adversarial linear bandit algorithm that enjoys a worst-case guarantee B_n on its n -round expected regret R_n when the adversary is restricted to playing $(\theta_t)_t$ in the polar \mathcal{A}° of \mathcal{A} . Show that if this algorithm is used in a stochastic linear bandit problem with parameter noise (that is, $\theta_t \sim \nu$) and the support of ν is a subset of the polar of \mathcal{A}° then the expected

regret R'_n is still bounded by B_n . Derive a bound on the expected regret R'_n in the stochastic problem when the restriction on ν is replaced by an assumption that $\sup_{a \in \mathcal{A}} \langle a, \theta_t - \theta \rangle$ has a bounded magnitude.

29.3 Modify LinUCB to make it (potentially) better for stochastic bandits under parameter noise. Show that the regret improves.

29.4 Complete the details to prove the claims made in Note 1. In particular, prove that for $\mathcal{A} = B_2^d$ there exists a universal constant $C > 0$ such that the expected regret R_n of an appropriately tuned version of the bandit algorithm of Theorem 28.4 satisfies $R_n \leq C(d\sqrt{n} + \varepsilon n)$, where $\varepsilon = \sup_{a \in \mathcal{A}} \varepsilon(a)$.

29.5 Complete the details to prove the claims made in Note 4.



You will need to repeat the analysis in Eq. (28.11), update the learning rate and check the bounds on the norm of the estimators.

Part VII

Other Topics

In the penultimate part we collect a few topics to which we could not dedicate a whole part. When deciding what to include we balanced our subjective views on what is important, pedagogical and sufficiently well-understood for a book. Of course we have played favourites with our choices and hope the reader can forgive us for the omissions. We spend the rest of this intro outlining some of the omitted topics.

Continuous-armed bandits

There is a small literature on bandits where the number of actions is infinitely large. We covered the linear case in earlier chapters, but the linear assumption can be relaxed significantly. Let \mathcal{A} be an arbitrary set and \mathcal{F} a set of functions from $\mathcal{A} \rightarrow \mathbb{R}$. The learner is given access to the action set \mathcal{A} and function class \mathcal{F} . In each round the learner chooses an action $A_t \in \mathcal{A}$ and receives reward $X_t = f(A_t) + \eta_t$ where η_t is noise and $f \in \mathcal{F}$ is fixed, but unknown. Of course this setup is general enough to model all of the stochastic bandits so far, but is perhaps too general to say much. One interesting relaxation is the case where \mathcal{A} is a metric space and \mathcal{F} is the set of Lipschitz functions. We refer the reader to papers by Kleinberg [2005], Auer et al. [2007], Kleinberg et al. [2008], Bubeck et al. [2011], Slivkins [2014].

Duelling bandits

In the duelling bandit problem the learner chooses two arms in each round A_{t1}, A_{t2} . Rather than observing a reward for each arm the learner observes the winner of a ‘dual’ between the two arms. Let K be the number of arms and $P \in [0, 1]^{K \times K}$ be a matrix where P_{ij} is the probability that arm i beats arm j in a dual. It is natural to assume that $P_{ij} = 1 - P_{ji}$. A common, but slightly less justifiable, assumption is the existence of a total ordering on the arms such that if $i \succ j$, then $P_{ij} > 1/2$. There are at least two notions of regret. Let i^* be the optimal arm so that $i^* \succ j$ for all $j \neq i^*$. Then the strong/weak regret are defined by

$$\text{Strong regret} = \mathbb{E} \left[\sum_{t=1}^n (P_{i^*, A_{t1}} + P_{i^*, A_{t2}} - 1) \right].$$

$$\text{Weak regret} = \mathbb{E} \left[\sum_{t=1}^n \min \{P_{i^*, A_{t1}} - 1/2, P_{i^*, A_{t2}} - 1/2\} \right].$$

Both definitions measure the number of times arms with low probability of winning a duel against the optimal arm is played. The former definition only vanishes when $A_{t1} = A_{t2} = i^*$, while the latter is zero as soon as $i^* \in \{A_{t1}, A_{t2}\}$. The duelling bandit problem was introduced by Yue et al. [2009] and has seen quite a lot of interest since then [Yue and Joachims, 2009, 2011, Ailon et al., 2014, Zoghi et al., 2014, Jamieson et al., 2015, Zoghi et al., 2015, Dudík et al., 2015, Komiya et al., 2015a, Wu and Liu, 2016].

Convex bandits

Let $\mathcal{A} \subset \mathbb{R}^d$ be a convex set. the convex bandit problem comes in both stochastic and adversarial varieties. In both cases the learner chooses A_t from \mathcal{A} . In the stochastic case the learner receives a reward $X_t = f(A_t) + \eta_t$ where f is an unknown convex function and η_t is noise. In the adversarial setting the adversary chooses a sequence of convex functions f_1, \dots, f_n and the learner receives reward $X_t = f(A_t)$. This turned out to be a major challenge over the last decade with most approaches leading to suboptimal regret in terms of the horizon. The best bounds in the stochastic case are by [Agarwal et al. \[2011\]](#) while in the adversarial case there has been a lot of recent progress [[Bubeck et al., 2015a](#), [Bubeck and Eldan, 2016](#), [Bubeck et al., 2017](#)]. In both cases the dependence of the regret on the horizon is $O(\sqrt{n})$, which is optimal in the worst case. Many open questions remain.

Budgeted bandits

In many problems choosing an action costs some resources. In the bandits with knapsacks problem the learner starts with a fixed budget $B \in [0, \infty)^d$ over d resource types. Like in the standard K -armed stochastic bandit, the learner chooses $A_t \in [K]$ and receives a reward X_t sampled from distribution depending on A_t . The twist is that the game does not end after a fixed number of rounds. Instead, in each round the environment samples a cost vector $C_t \in [0, 1]^d$ from a distribution that depends on A_t . The game ends in the first round τ where there exists an $i \in [d]$ such that $\sum_{t=1}^{\tau} C_{ti} > B_i$. This line of work was started by [Badanidiyuru et al. \[2013\]](#) and has been extended in many directions by [Agrawal and Devanur \[2014\]](#), [Tran-Thanh et al. \[2012\]](#), [Ashwinkumar et al. \[2014\]](#), [Xia et al. \[2015\]](#), [Agrawal and Devanur \[2016\]](#), [Tran-Thanh et al. \[2010\]](#), [Hanawal et al. \[2015\]](#). A somewhat related idea is the conservative bandit problem where the goal is to minimize regret subject to the constraint that the learner must not be much worse than some known baseline. The constraint limits the amount of exploration and makes the regret guarantees slightly worse [[Sui et al., 2015](#), [Wu et al., 2016](#), [Kazerouni et al., 2017](#)].

Learning with delays

In many practical applications the feedback to the learner is not immediate. The time between clicking on a link and buying a product could be minutes, days, weeks or longer. Similarly, the response to a drug does not come immediately. In most cases the learner does not have the choice to wait before making the next decision. Buyers and patients just keep coming. Perhaps the first paper for online learning with delays is by [Weinberger and Ordentlich \[2002\]](#), who consider the full information setting. Recently this has become a hot topic and there has been a lot of follow-up work extending the results in various directions [[Joulani et al., 2013](#), [Desautels et al., 2014](#), [Cesa-Bianchi et al., 2016](#), [Vernade et al., 2017, 2018](#), [Pike-Burke et al., 2018](#), and others]. Learning with delays is an interesting example where the adversarial and stochastic models lead to quite

different outcomes. In general the increase in regret due to rewards being delayed by at most τ rounds is a multiplicative $\sqrt{\tau}$ factor for adversarial models and an additive term only for stochastic models.

Graph feedback

There is growing interest in feedback models that lie between the full information and bandit settings. One way to do this is to let G be a directed graph with K vertices. The adversary chooses a sequence of loss vectors in $[0, 1]^K$ as usual. In each round the learner chooses a vertex and observes the loss corresponding to that vertex and its neighbours. The full information and bandit settings are recovered by choosing the graph to be fully connected or have no edges respectively, but of course there are many interesting regimes in between. There are many variants on this basic problem. For example, G might change in each round or be undirected. Or perhaps the graph is changing and the learner only observes it after choosing an action. The reader can explore this topic by reading the articles by [Mannor and Shamir \[2011\]](#), [Alon et al. \[2013\]](#), [Kocák et al. \[2014\]](#), [Alon et al. \[2015\]](#) or the short book by [Valko \[2016\]](#).

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

30 Combinatorial Bandits

A combinatorial bandit is a linear bandit with a special kind of combinatorial action-set $\mathcal{A} \subset \{0, 1\}^d$, which for constant $m \in [d]$ satisfies

$$\mathcal{A} \subseteq \{a \in \{0, 1\}^d : \|a\|_1 \leq m\}.$$

The setting is studied in both adversarial and stochastic models. We focus on the former in this chapter and discuss the latter in the notes. As usual the adversary chooses a sequence of loss vectors y_1, \dots, y_n with $y_t \in \mathbb{R}^d$ and the expected regret is

$$R_n = \mathbb{E} \left[\max_{a \in \mathcal{A}} \sum_{t=1}^n \langle A_t - a, y_t \rangle \right].$$

In Chapters 27 and 28 we assumed that $y_t \in \mathcal{A}^\circ$, which guarantees that $|\langle A_t, y_t \rangle| \leq 1$ for all t . This restriction is not consistent with the applications we have in mind, so instead we assume that $y_t \in [0, 1]^d$, which by the definition of \mathcal{A} ensures that $|\langle A_t, y_t \rangle| \leq m$ for all t . In the standard bandit model the learner observes $\langle A_t, y_t \rangle$ in each round. In many applications for which combinatorial bandits are applied there is actually more information available. The simplest is the **full information** setting where the learner observes the whole vector y_t . The full information setup is interesting, but does not have the flavor of a bandit problem and so we do not discuss it further. There is an inbetween setting where the learner receives **semibandit** feedback, which is the vector $(A_{t1}y_{t1}, \dots, A_{td}y_{td})$. Since $A_{ti} \in \{0, 1\}$ this is equivalent to observing y_{ti} for those i when $A_{ti} = 1$.

30.1 Applications

Shortest path problems

The online shortest path problem is a game over n between adversary and learner. Let $G = (V, E)$ be a fixed graph with a finite set of vertices V and edges $E \subseteq V \times V$. At the beginning of the game the adversary chooses the length of each edge in an arbitrary way. In each round the learner chooses a path between fixed vertices $u, v \in V$ with the goal of travelling the shortest distance. The regret of the learner is the difference between the distance they travelled and

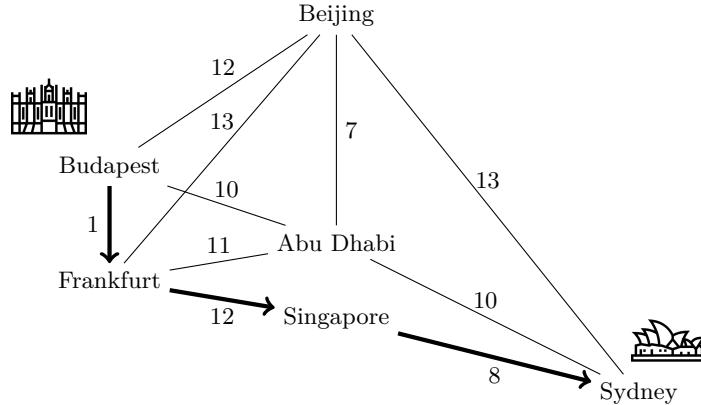


Figure 30.1 Shortest-path problem between Budapest and Sydney. The learner chooses the path Budapest–Frankfurt–Singapore–Sydney. In the bandit setting they observe total travel time (21 hours) while in the semibandit they observe the length of each flight on the route they took (1 hour, 12 hours, 8 hours).

the distance of the optimal path in hindsight. To make things a little formal, let $d = |E|$ and for $t \in [n]$ and $i \in [d]$ let $y_{ti} \in [0, 1]$ be the length of the i th path in round t as chosen by the adversary. A path is represented by a vector $a \in \{0, 1\}^d$ where $a_i = 1$ if the i th edge is part of the path. Let \mathcal{A} be the set of paths connecting vertices u and v , then the length of path a in round t is $\langle a, y_t \rangle$.

Ranking

Suppose a company has d possible ads they can place and m locations in which to display them. In each round t the learner should choose the m ads to display, which is represented by a vector $A_t \in \{0, 1\}^d$ with $\|A_t\|_1 = m$. As before, the adversary chooses $y_t \in [0, 1]^d$ that measures the quality of each placement and the learner suffers loss $\langle A_t, y_t \rangle$. This problem could also be called ‘selection’ because there is no ordering. This is not true in applications like web search where the order of search results is as important as the results themselves. This kind of problem is analyzed in Chapter 32.

Multitask bandits

Consider playing m multi-armed bandits simultaneously, each with K arms. If the losses for each bandit problem are observed, then it is easy to apply Exp3 or Exp3-IX to each bandit independently. But now suppose the learner only observes the sum of the losses. This problem is represented as a combinatorial bandit by letting $d = mK$ and

$$\mathcal{A} = \left\{ a \in \{0, 1\}^d : \sum_{i=1}^K a_{i+Kj} = 1 \text{ for all } 0 \leq j < m \right\}.$$

This scenario can arise in practice when a company is making multiple independent interventions, but the quality of the interventions are only observed via a change in revenue.

30.2 Bandits

The easiest approach is to apply Exp3 with John's exploration as described in Chapter 27. The only difference is that now $|\langle A_t, y_t \rangle|$ can be as large as m , which increases the regret by a factor of m . We leave the proof of the following theorem to the reader (Exercise 30.1).

THEOREM 30.1 *If Algorithm 14 is run on action-set \mathcal{A} with appropriately chosen learning rate, then*

$$R_n \leq 2m\sqrt{3dn \log |\mathcal{A}|} \leq m^{3/2} \sqrt{12dn \log \left(\frac{ed}{m}\right)}.$$

There are two computational issues with this approach. First, the action-set is typically so large that finding the core set of the central minimum volume enclosing ellipsoid that determines the exploration distribution of Algorithm 14 is hopeless. Second, sampling from the resulting exponential weights distribution may be highly nontrivial. There is no golden bullet for these issues. We cannot expect the travelling salesman to get easier when done online and with bandit feedback. There are, however, some special cases where efficient algorithms exist and we give some pointers to the literature at the end of the chapter. One modification that greatly eases computation is to replace John's exploration with something more tractable. Let $\pi : \mathcal{A} \rightarrow [0, 1]$ be the exploration distribution used by Algorithm 14 and $Q(\pi) = \sum_{a \in \mathcal{A}} \pi(a)aa^\top$. Then the regret of Algorithm 14 satisfies

$$R_n = O \left(m \sqrt{\max_{a \in \mathcal{A}} \|a\|_{Q(\pi)^{-1}}^2 n \log(|\mathcal{A}|)} \right).$$

By Kiefer–Wolfowitz (Theorem 21.1) we know that π can be chosen so that $\|a\|_{Q(\pi)^{-1}}^2 = d$ and that if $\text{span}(\mathcal{A}) = \mathbb{R}^d$, then this is optimal. In many cases, however, a similar result can be proven for other exploration distributions with more attractive properties computationally.

30.3 Semibandits

The additional information is easily exploited by noting that y_t can now be estimated in each coordinate. Let

$$\hat{Y}_{ti} = \frac{A_{ti}y_{ti}}{P_{ti}}, \quad (30.1)$$

where $P_{ti} = \mathbb{E}[A_{ti} | \mathcal{F}_{t-1}]$ with $\mathcal{F}_t = \sigma(A_1, Z_1, \dots, A_t, Z_t)$. An easy calculation shows that $\mathbb{E}[\hat{Y}_{ti} | \mathcal{F}_{t-1}] = y_{ti}$.

```

1: Input  $\mathcal{A}, \eta, F$ 
2:  $\bar{A}_1 = \operatorname{argmin}_{a \in \mathcal{A}} F(a)$ 
3: for  $t = 1, \dots, n$  do
4:   Choose  $P_t$  on  $\mathcal{A}$  such that  $\sum_{a \in \mathcal{A}} P_t(a)a = \bar{A}_t$ 
5:   Sample  $A_t \sim P_t$ 
6:   Compute  $\hat{Y}_{ti} = \frac{A_{ti}y_{ti}}{P_{ti}}$  for all  $i \in [d]$ 
7:   Update  $\bar{A}_{t+1} = \operatorname{argmin}_{a \in \text{co}(\mathcal{A})} \eta \langle a, \hat{Y}_t \rangle + D_F(a, \bar{A}_t)$ 
8: end for
```

Algorithm 16: Online stochastic mirror descent for semibandits

THEOREM 30.2 Let $F : \mathbb{R}^d \rightarrow \mathbb{R}$ be the negentropy potential defined by

$$F(a) = \sum_{i=1}^d (a_i \log(a_i) - a_i) .$$

If Algorithm 16 is run with $\eta = \sqrt{2m(1 + \log(d/m))/(nd)}$, then

$$R_n \leq \sqrt{2nmd(1 + \log(d/m))} .$$

Proof Recall from Chapter 28 that for Legendre potentials the optimization problem for \bar{A}_{t+1} can be written in a two-step process:

$$\begin{aligned} \nabla F(\tilde{A}_{t+1}) &= \nabla F(\bar{A}_t) - \eta \hat{Y}_t \\ \bar{A}_{t+1} &= \operatorname{argmin}_{a \in \text{co}(\mathcal{A})} D_F(a, \tilde{A}_{t+1}) . \end{aligned}$$

Then by Theorem 28.3 we have

$$R_n \leq \frac{\operatorname{diam}_F(\text{co}(\mathcal{A}))}{\eta} + \frac{1}{\eta} \sum_{t=1}^n \mathbb{E} [D_{F^*}(\nabla F(\tilde{A}_{t+1}), \nabla F(\bar{A}_t))] .$$

The Legendre-Fenchel dual is $F^*(u) = \sum_{i=1}^d \exp(u_i)$ and the Bregman divergence with respect to this potential is

$$D_{F^*}(u, v) = \sum_{i=1}^d (\exp(u_i) - \exp(v_i)) - \sum_{i=1}^d (u_i - v_i) \exp(v_i) .$$

Since $\nabla F(a)_i = \log(a_i)$ we have

$$\begin{aligned} D_{F^*}(\nabla F(\tilde{A}_{t+1}), \nabla F(\bar{A}_t)) &= \sum_{i=1}^d (\tilde{A}_{t+1,i} - \bar{A}_{ti}) + \sum_{i=1}^d \eta \bar{A}_{ti} \hat{Y}_{ti} \\ &= \sum_{i=1}^d \bar{A}_{ti} \left(\exp(-\eta \hat{Y}_{ti}) - 1 + \eta \hat{Y}_{ti} \right) \leq \frac{\eta^2}{2} \sum_{i=1}^d \bar{A}_{ti} \hat{Y}_{ti}^2 . \end{aligned}$$

where the inequality follows from the fact that $\exp(-x) \leq 1 - x + x^2/2$ for all $x \geq 0$. Taking the expectation leads to

$$\mathbb{E} \left[\sum_{i=1}^d \bar{A}_{ti} \hat{Y}_{ti}^2 \right] = \mathbb{E} \left[\sum_{i=1}^d \frac{y_{ti}^2 A_{ti}}{\bar{A}_{ti}} \right] \leq d.$$

The diameter is easily bounded by noting that F is negative in $\text{co}(\mathcal{A})$ and using the Cauchy-Schwartz inequality:

$$\begin{aligned} \text{diam}_F(\text{co}(\mathcal{A})) &= \sup_{a \in \text{co}(\mathcal{A})} \sum_{i=1}^d \left(a_i \log(a_i) - a_i + \bar{A}_{1i} + \bar{A}_{1i} \log\left(\frac{1}{\bar{A}_{1i}}\right) \right) \\ &\leq m + \sum_{i=1}^d \bar{A}_{1i} \log\left(\frac{1}{\bar{A}_{1i}}\right) \leq m(1 + \log(d/m)). \end{aligned}$$

Putting together the pieces shows that

$$R_n \leq \frac{m(1 + \log(d/m))}{\eta} + \frac{\eta nd}{2} = \sqrt{2nmd(1 + \log(d/m))}. \quad \square$$



Algorithm 16 plays mirror descent on the convex hull of the actions, which has dimension $d - 1$. In principle it would be possible to do the same thing on the set of distributions over actions, which has dimension K . Repeating the analysis leads to a suboptimal regret of $O(m\sqrt{dn \log(d/m)})$. We encourage the reader to go through this calculation to see where things go wrong.

Like in Section 30.2, the main problem is computation. There are two challenges: First, in each round the algorithm needs to find a distribution P_t over \mathcal{A} such that $\sum_{a \in \mathcal{A}} P_t(a) = \bar{A}_t$. Feasibility follows from the definition of $\text{co}(\mathcal{A})$ while Carathéodory's theorem proves the support of P_t never needs to be larger than $d + 1$. Since \mathcal{A} is finite we can write this problem in terms of linear constraints, but naively the computation complexity is polynomial in K , which is exponential in m . The second difficulty is computing \bar{A}_{t+1} from \bar{A}_t and \hat{Y}_t . This is a convex optimization problem, but the computation complexity depends on the representation of \mathcal{A} and may be intractable.

30.4 Follow the perturbed leader

The computational complexity of mirror descent in the previous section can be prohibitively expensive. In this section we describe an efficient algorithm for online combinatorial optimization under the assumption that for all $y \in [0, 1]^d$ the optimization problem of finding

$$a = \operatorname{argmin}_{a \in \mathcal{A}} \langle a, y \rangle \tag{30.2}$$

admits an efficient solution. This feels like the minimum one could get away with. If the static problem is too hard it seems unlikely that an online algorithm could be efficient. In fact, an online algorithm with low regret could be used to approximate the solution to the static problem.

So we will try to design an algorithm for which the only nontrivial computation is solving Eq. (30.2). The **follow-the-perturbed leader** (FTPL) algorithm operates by estimating the cumulative losses observed so far. In each round the estimates are perturbed by some random amount and the algorithm solves Eq. (30.2) using the perturbed estimates. Let $\hat{L}_t = \sum_{s=1}^t \hat{Y}_t$ be the cumulative loss estimates after round t , then FTPL chooses

$$A_{t+1} = \operatorname{argmin}_{a \in \mathcal{A}} \langle a, \eta \hat{L}_t - Z_{t+1} \rangle, \quad (30.3)$$

where $\eta > 0$ is the learning rate and $Z_{t+1} \in \mathbb{R}^d$ is a random perturbation sampled from distribution Q to be chosen later. The random perturbations introduce the exploration, which if for appropriate perturbation distributions is sufficient to guarantee small regret. Notice that if η is small, then the effect of Z_{t+1} is larger and the algorithm can be expected to explore more, which is consistent with the learning rate used in mirror descent or exponential weights studied in previous chapters.

We still need to define the loss estimates and perturbation distribution. First we make a connection between this algorithm and mirror descent. Given Legendre potential F with $\operatorname{dom}(\nabla F) = \operatorname{int}(\mathcal{A})$ online stochastic mirror descent chooses \bar{A}_{t+1} so that

$$\bar{A}_{t+1} = \operatorname{argmin}_{a \in \mathcal{A}} \langle a, \eta \hat{Y}_t \rangle + D_F(a, \bar{A}_t).$$

Taking derivatives and using the fact that $\operatorname{dom}(\nabla F) = \operatorname{int}(\mathcal{A})$ we have

$$\nabla F(\bar{A}_{t+1}) = \nabla F(\bar{A}_t) - \eta \hat{Y}_t = -\eta \hat{L}_t.$$

By duality this implies that $\bar{A}_{t+1} = \nabla F^*(-\eta \hat{L}_t)$ where $F^*(x) = \sup_{a \in \mathcal{A}} (\langle x, a \rangle - F(a))$ is the Fenchel conjugate of F . Examining Eq. (30.3) we see that

$$\bar{A}_{t+1} = \mathbb{E}[A_{t+1} | \mathcal{F}_t] = \mathbb{E} \left[\operatorname{argmin}_{a \in \mathcal{A}} \langle a, \eta \hat{L}_t - Z_{t+1} \rangle \right].$$

If we are to view follow-the-perturbed leader as an instance of mirror descent we must find a Legendre potential F with

$$\nabla F^*(-\eta \hat{L}_t) = \mathbb{E} \left[\operatorname{argmin}_{a \in \mathcal{A}} \langle a, \eta \hat{L}_t - Z \rangle \right] = \mathbb{E} \left[\operatorname{argmax}_{a \in \mathcal{A}} \langle a, Z - \eta \hat{L}_t \rangle \right],$$

which is equivalent to $\nabla F^*(x) = \mathbb{E}[\operatorname{argmax}_{a \in \mathcal{A}} \langle a, x + Z \rangle]$. In order to remove clutter in the notation we define

$$a(x) = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, x \rangle.$$

Readers with some familiarity with convex analysis will remember that if $\phi(x) = \max_{a \in \mathcal{A}} \langle a, x \rangle$ is the support function and \mathcal{A} has a smooth boundary, then $\nabla \phi(x) = a(x)$. For combinatorial bandits \mathcal{A} is not smooth, but if Q is

absolutely continuous with respect to the Lebesgue measure, then you will show in Exercise 30.3 that nevertheless it is true that

$$\nabla \mathbb{E} [\phi(x + Z)] = \mathbb{E} [a(x + Z)].$$

All this shows that follow-the-perturbed-leader can be interpreted as mirror descent with potential F defined in terms of its Fenchel dual.

$$F^*(x) = \mathbb{E} [\phi(x + Z)]. \quad (30.4)$$

There are more reasons for making this connection than mere curiosity. The classical analysis of FTPL is highly probabilistic and involves at least one ‘leap of faith’ in the analysis. In contrast, the analysis via the mirror descent interpretation is more mechanical. Recall that mirror descent depends on choosing a potential, an exploration distribution and an estimator. The exploration distribution is a distribution P_t on \mathcal{A} such that

$$\bar{A}_t = \sum_{a \in \mathcal{A}} P_t(a)a,$$

which in our case is simply defined by

$$P_t(a) = \mathbb{P}(a(Z - \eta \hat{L}_{t-1}) = a \mid \mathcal{F}_{t-1}).$$

It remains to choose the loss estimator. A natural choice would be the same as Eq. (30.1), which is

$$\hat{Y}_{ti} = \frac{A_{ti}y_{ti}}{P_{ti}},$$

where $P_{ti} = \mathbb{P}(A_{ti} = 1 \mid \mathcal{F}_{t-1})$. The problem is that

$$P_{ti} = \mathbb{P}(a(Z - \eta \hat{L}_{t-1})_i = 1 \mid \mathcal{F}_{t-1}),$$

which does not generally have a closed form solution. If computation were not an issue, then we could simply estimate P_{ti} for each i by sampling. The trick is to notice that we actually only need to estimate the reciprocal $1/P_{ti}$. Let $X \in \{1, 2, \dots\}$ be a geometrically distributed random variable with parameter $\theta \in [0, 1]$ so that

$$\mathbb{P}(X = k) = (1 - \theta)^{k-1}\theta^k.$$

An easy calculation shows that $\mathbb{E}[X] = 1/\theta$. Let $K_{it} \in \{1, 2, \dots\}$ be chosen so that $\mathbb{P}(K_{it} = k \mid \mathcal{F}_{t-1}) = (1 - P_{it})^{k-1}P_{it}$. Then for constant $\beta > 0$ define

$$\hat{Y}_{ti} = \min\{\beta, K_{it}A_{it}y_{it}\}.$$

The truncation parameter β is needed to ensure that \hat{Y}_{ti} is never too large, but note that without we have

$$\mathbb{E}_{t-1}[K_{it}A_{it}y_{it}] = y_{it}.$$

We have now provided all the pieces to define mirror descent. The algorithm is summarized in Algorithm 17.



In Chapter 28 we assumed the loss estimator was unbiased, but this is not necessary as we shall see in the analysis.

```

1: Input  $\mathcal{A}, n, \eta, \beta, Q$ 
2:  $\hat{L}_{0i} = 0$  for each  $i \in [d]$ 
3: for  $t = 1, \dots, n$  do
4:   Sample  $Z_t \sim Q$ 
5:   Compute  $A_t = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \eta \hat{L}_{t-1} - Z_t \rangle$ 
6:   For each  $i$  with  $A_{ti} = 1$  sample  $K_{ti} \sim \text{Geometric}(P_{ti})$ 
7:    $\hat{Y}_{ti} = \min \{\beta, A_{ti} K_{ti} y_{ti}\}$ 
8:    $\hat{L}_{ti} = \hat{L}_{t-1,i} + \hat{Y}_{ti}$ 
9: end for
```

Algorithm 17: Follow-the-Perturbed leader for semibandits

THEOREM 30.3 Let Q have density $q(z) = 2^{-d} \exp(-\|z\|_1)$ and

$$\eta = \sqrt{\frac{1 + \log(d)}{nd}} \quad \beta = \frac{1}{\eta m}.$$

Then the regret of Algorithm 17 is bounded by $R_n \leq 2(m \vee e)\sqrt{nd(1 + \log(d))}$.

Proof First we subtract the bias in the loss estimators and apply Theorem 28.1 to show that

$$\begin{aligned}
R_n(a) &= \mathbb{E} \left[\sum_{t=1}^n \langle A_t - a, y_t \rangle \right] = \mathbb{E} \left[\sum_{t=1}^n \langle \bar{A}_t - a, y_t \rangle \right] \\
&= \mathbb{E} \left[\sum_{t=1}^n \langle \bar{A}_t - a, \hat{Y}_t \rangle \right] + \mathbb{E} \left[\sum_{t=1}^n \langle \bar{A}_t - a, y_t - \hat{Y}_t \rangle \right] \\
&\leq \frac{\operatorname{diam}_F(\mathcal{A})}{\eta} + \mathbb{E} \left[\frac{1}{\eta} \sum_{t=1}^n D_F(\bar{A}_t, \bar{A}_{t+1}) \right] + \mathbb{E} \left[\sum_{t=1}^n \langle \bar{A}_t - a, y_t - \hat{Y}_t \rangle \right]. \quad (30.5)
\end{aligned}$$

Of the three terms the diameter is most easily bounded.

$$\begin{aligned}
F(a) &= \sup_{x \in \mathbb{R}^d} (\langle x, a \rangle - F^*(x)) = \sup_{x \in \mathbb{R}^d} (\langle x, a \rangle - \mathbb{E}[\max_{b \in \mathcal{A}} \langle x + Z, b \rangle]) \quad (30.6) \\
&\geq -\mathbb{E}[\max_{b \in \mathcal{A}} \langle Z, b \rangle] \geq -m\mathbb{E}[\|Z\|_\infty] = -m \sum_{i=1}^d \frac{1}{d} \geq -m(1 + \log(d)),
\end{aligned}$$

where the first inequality follows by choosing $x = 0$ and the second follows from Holder's inequality. The last equality is nontrivial and is explained in Exercise 30.2. By the convexity of the maximum function and the fact that Z is centered we also have from Eq. (30.6) that $F(a) \leq 0$, which means that

$$\operatorname{diam}_F(\mathcal{A}) = \max_{a, b \in \mathcal{A}} F(a) - F(b) \leq m(1 + \log(d)). \quad (30.7)$$

The next step is to bound the Bregman divergence induced by F . We will shortly show that the Hessian $\nabla^2 F^*(x)$ of F^* exists, so by duality and Taylor's theorem there exists an $\alpha \in [0, 1]$ and $\xi = -\eta \hat{L}_{t-1} - \alpha \eta \hat{Y}_t$ such that

$$\begin{aligned} D_F(\bar{A}_t, \bar{A}_{t+1}) &= D_{F^*}(\nabla F(\bar{A}_{t+1}), \nabla F(\bar{A}_t)) \\ &= D_{F^*}(-\eta \hat{L}_{t-1} - \eta \hat{Y}_t, \nabla F(-\eta \hat{L}_{t-1})) = \frac{\eta^2}{2} \|\hat{Y}_t\|_{\nabla^2 F^*(\xi)}^2, \end{aligned} \quad (30.8)$$

where the last equality follows from Taylor's theorem (see Theorem 26.4). To calculate the Hessian we use a change of variable to avoid applying the gradient to the non-differentiable argmax.

$$\begin{aligned} \nabla^2 F^*(x) &= \nabla(\nabla F(x)) = \nabla \mathbb{E}[a(x + Z)] = \nabla \int_{\mathbb{R}^d} a(x + z) f(z) dz \\ &= \nabla \int_{\mathbb{R}^d} a(u) f(u - x) du = \int_{\mathbb{R}^d} a(u) (\nabla f(u - x))^T du \\ &= \int_{\mathbb{R}^d} a(u) \text{sign}(u - x)^T f(u - x) du = \int_{\mathbb{R}^d} a(x + z) \text{sign}(z)^T f(z) dz. \end{aligned}$$

Using the definition of ξ and the fact that $a(x)$ is nonnegative,

$$\begin{aligned} \nabla^2 F^*(\xi)_{ij} &= \int_{\mathbb{R}^d} a(\xi + z)_i \text{sign}(z)_j f(z) dz \end{aligned} \quad (30.9)$$

$$\begin{aligned} &\leq \int_{\mathbb{R}^d} a(\xi + z)_i f(z) dz \\ &= \int_{\mathbb{R}^d} a(z - \eta \hat{L}_{t-1} - \alpha \eta \hat{Y}_t)_i f(z) dz \\ &= \int_{\mathbb{R}^d} a(u - \eta \hat{L}_{t-1})_i f(u + \alpha \eta \hat{Y}_t) du \\ &\leq \exp\left(\|\alpha \eta \hat{Y}_t\|_1\right) \int_{\mathbb{R}^d} a(u - \eta \hat{L}_{t-1})_i f(u) du \\ &\leq e P_{ti}, \end{aligned} \quad (30.10)$$

where the last inequality follows since $\alpha \in [0, 1]$ and $\hat{Y}_{ti} \leq \beta = 1/(m\eta)$ and \hat{Y}_t has at most m nonzero entries. Continuing on from Eq. (30.8) we have

$$\frac{\eta^2}{2} \|\hat{Y}_t\|_{\nabla^2 F^*(\xi)}^2 \leq \frac{e\eta^2}{2} \sum_{i=1}^d P_{ti} \hat{Y}_{ti} \sum_{j=1}^d \hat{Y}_{tj} \leq \frac{e\eta}{2} \sum_{i=1}^d P_{ti} \hat{Y}_{ti} \leq \frac{e\eta}{2} \sum_{i=1}^d P_{ti} K_{ti}.$$

Chaining together the parts and taking the expectation shows that

$$\mathbb{E}[D_F(\bar{A}_t, \bar{A}_{t+1})] \leq \frac{e\eta}{2} \mathbb{E}\left[\sum_{i=1}^d P_{ti} K_{ti}\right] = \frac{e\eta}{2} \mathbb{E}\left[\sum_{i=1}^d P_{ti} \mathbb{E}[K_{ti} \mid \mathcal{F}_{t-1}]\right] = \frac{ed\eta}{2}.$$

The last step is to control the bias term.

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^n \langle \bar{A}_t - a, y_t - \hat{Y}_t \rangle \right] &\leq \mathbb{E} \left[\sum_{t=1}^n \langle \bar{A}_t, y_t - \hat{Y}_t \rangle \right] \\ &= \mathbb{E} \left[\sum_{t=1}^n \sum_{i=1}^d P_{ti} (y_{ti} - \min \{P_{ti}\beta, y_{ti}\}) \right] \leq \frac{dn}{2\beta} = \frac{dnm\eta}{2}. \end{aligned}$$

Putting together all the pieces into Eq. (30.5) leads to

$$R_n \leq \frac{m(1 + \log(d))}{\eta} + \frac{end\eta}{2} + \frac{dnm\eta}{2} \leq 2(m \vee e)\sqrt{nd(1 + \log(d))}. \quad \square$$

30.5 Notes

- 1 For a long time it was speculated that the dependence of the regret on $m^{3/2}$ in Theorem 30.1 might be improvable to m . Very recently, however, the lower bound was increased to show the upper bound is tight [Cohen et al., 2017]. For semibandits the worst case lower bound is $\Omega(\sqrt{dnm})$ (Exercise 30.5), which is matched up to constant factors by online stochastic mirror descent with a different potential (Exercise 30.4).
- 2 Algorithm 17 needs to sample K_{ti} for each i with $A_{ti} = 1$. The conditional expected running time for this is A_{ti}/P_{ti} , which has expectation 1. It follows that the expected running time over the whole n rounds is $O(nd)$ calls to the oracle linear optimization algorithm. It can happen that the algorithm is unlucky and chooses $A_{ti} = 1$ for some i with P_{ti} quite small. To avoid catastrophic slowdowns it is possible to truncate the sampling procedure by defining $\tilde{K}_{ti} = \min\{K_{ti}, M\}$ for M suitably large. This introduces a small controllable bias [Neu, 2015a].
- 3 While FTRL is excellent in the face of semibandit information, we do not know of a general result for the bandit model. The main challenge is controlling the variance of the least squares estimator without John's exploration.
- 4 Combinatorial bandits can also be studied in a stochastic setting. There are several ways to do this. The first mirrors our assumptions for stochastic linear bandits in Chapter 19 where the loss (more commonly reward) is defined by

$$X_t = \langle A_t, \theta \rangle + \eta_t, \quad (30.11)$$

where $\theta \in \mathbb{R}^d$ is fixed and unknown and η_t is the noise on which statistical assumptions are made (for example, conditionally 1-subgaussian). There are at least two alternatives. Suppose that $\theta_1, \dots, \theta_n$ are sampled independently from some multivariate distribution and define the reward by

$$X_t = \langle A_t, \theta_t \rangle. \quad (30.12)$$

This latter version has ‘parameter noise’ and is more closely related to the adversarial setup studied in this chapter. Finally, one can assume additionally

that the distribution of θ_t is a product distribution so that $(\theta_{1i})_{i=1}^d$ are also independent.

- 5 For some action-sets the off-diagonal elements of the Hessian in Eq. (30.9) are negative, which improves the dependence on m to just \sqrt{m} . An example where this occurs is when $\mathcal{A} = \{a \in \{0, 1\}^d : \|a\|_1 = m\}$. Let $i \neq j$ and suppose that $z, \xi \in \mathbb{R}^d$ and $z_j \geq 0$. Then you can check that $a(z + \xi)_i \leq a(z - 2z_j e_j + \xi)_i$ and so

$$\begin{aligned}\nabla^2 F^*(\xi)_{ij} &= \int_{\mathbb{R}^d} a(z + \xi)_i \operatorname{sign}(z)_j f(z) dz \\ &= \int_{\mathbb{R}^{d-1}} \int_0^\infty (a(z + \xi)_i - a(z - 2z_j e_j + \xi)_i) f(z) dz_j dz_{-j} \\ &\leq 0,\end{aligned}$$

where dz_{-j} is shorthand for $dz_1 dz_2, \dots, dz_{j-1} dz_{j+1}, \dots, dz_d$. You are asked to complete all the details in Exercise 30.6. This result unfortunately does not hold for every action-set (Exercise 30.7).

30.6 Bibliographic remarks

The online combinatorial bandit was introduced by Cesa-Bianchi and Lugosi [2012] where the most comprehensive list of known applications for which computation is efficient is given. The analysis presented in Section 30.2 is due to Bubeck and Cesa-Bianchi [2012]. While computational issues remain in the bandit problem, there has been some progress in certain settings [Combes et al., 2015]. The full information setting has been studied quite extensively [Koolen et al., 2010, and references from/to]. The follow-the-perturbed-leader algorithm was first proposed by Hannan [1957], rediscovered by Kalai and Vempala [2005a,b] and generalized by Poland [2005], Hutter and Poland [2005]. Poland [2005] showed a near-optimal regret for finite-armed adversarial bandits while for combinatorial settings suboptimal rates have been shown by Awerbuch and Kleinberg [2004], McMahan and Blum [2004]. Semibandits seem to have been introduced in the context of shortest-path problems by György et al. [2007]. The general setup and algorithmic analysis of FTPL presented follows the work by Neu [2015a] who also had the idea to estimate the inverse probabilities via a geometric random variable. Our analysis based on mirror descent improves the regret by a factor of \sqrt{m} . As far as we know this has not been seen in the literature on combinatorial bandits before, but the approach is heavily inspired by Abernethy et al. [2014] who present the core ideas in the prediction with expert advice setting, Cohen and Hazan [2015] in the combinatorial full information case and Abernethy et al. [2015] for finite-armed bandits. The literature on stochastic combinatorial semibandits is also quite large with algorithms and analysis in the frequentist [Gai et al., 2012, Combes et al., 2015, Kveton et al., 2015b] and Bayesian settings [Wen et al., 2015, Russo and Roy, 2016]. These works focus on the case where the reward

is given by Eq. (30.12) and the components of θ_t are independent. When the reward is given by Eq. (30.11) one can use the tools for stochastic linear bandits developed in Part V.

30.7 Exercises

30.1 Prove Theorem 30.1.

30.2 Let Z be sampled from measure on \mathbb{R}^d with density $f(z) = 2^{-d} \exp(-\|z\|_1)$. The purpose of this exercise is to show that

$$\mathbb{E}[\|Z\|_\infty] = \sum_{i=1}^d \frac{1}{i}. \quad (30.13)$$

Recall that an exponential with rate λ has density $f(x) = \lambda \exp(-\lambda x) \mathbb{1}_{x \geq 0}$.

- (a) Let X be exponential with rate λ . Show that $\mathbb{E}[X] = 1/\lambda$.
- (b) Let X_1, \dots, X_i be independent and exponentially distributed with rate 1. Show that $M = \min_{j \in [i]} X_j$ is exponentially distributed with rate i .
- (c) Show that $\|Z\|_\infty$ has the same law as the maximum of d independent standard exponentials.
- (d) Let M_1, \dots, M_d be independent exponentially distributed random variables where M_i has rate i . Show that Z has the same law as $\sum_{i=1}^d M_i$.
- (e) Show that Eq. (30.13) holds.

30.3 Let $\mathcal{A} \subset \mathbb{R}^d$ be a compact convex set and $\phi(x) = \max_{a \in \mathcal{A}} \langle a, x \rangle$ its support function. Let Q be a measure on \mathbb{R}^d that is absolutely continuous with respect to the Lebesgue measure and let $Z \sim Q$. Show that

$$\nabla \mathbb{E}[\phi(x + Z)] = \mathbb{E}[\operatorname{argmax}_{a \in \mathcal{A}} \langle a, x + Z \rangle].$$

30.4 Adapt the analysis in Exercise 28.10 to derive an algorithm for combinatorial bandits with semibandit feedback for which the regret is $R_n \leq C\sqrt{mdn}$ for universal constant $C > 0$.

30.5 Let $m \geq 1$ and that $d = km$ for some $k > 1$. Prove that for any algorithm there exists a combinatorial semibandit such that $R_n \geq c \min\{nm, \sqrt{mdn}\}$ where $c > 0$ is a universal constant.



The most obvious choice is to choose $\mathcal{A} = \{a \in \{0, 1\}^d : \|a\|_1 = m\}$, which are sometimes called m -sets. A lower bound does hold for this action-set [Lattimore et al., 2018]. However an easier path is to impose a little additional structure and analyze the multitask bandit setting.

30.6 Use the ideas in Note 5 to prove that FTPL has $R_n = \tilde{O}(\sqrt{mnd})$ regret when $\mathcal{A} = \{a \in \{0, 1\}^d : \|a\|_1 = m\}$.



After proving the off-diagonal elements of the Hessian are negative you will also need to tune the learning rate. We do not know of a source for this result, but the full information case was studied by Cohen and Hazan [2015].

30.7 Construct an action-set and $i \neq j$ and $z, \xi \in \mathbb{R}^d$ with $z_j > 0$ such that $a(z + \xi)_i \geq a(z - 2z_j e_j + \xi)_i$.

31 Non-Stationary Bandits

The usual definition of regret is not a suitable measure of performance when the underlying environment is changing. The purpose of this chapter is to provide a meaningful definition for nonstationary environments and show that algorithms exist for which this regret is sublinear. While we specify the results to bandits with finitely many arms (both stochastic and adversarial), many of the ideas generalize to other models such as linear bandits.

This chapter also illustrates the flexibility of the tools presented in the earlier chapters, which are applied here almost without modification. We hope (and expect) that this will also be true for other models you might study.

31.1 Adversarial bandits

In contrast to stochastic bandits, the adversarial bandit model presented in Chapter 11 does not prevent the environment from changing over time. The problem is that bounds on the regret can become vacuous when the losses appear nonstationary. To illustrate an extreme situation, suppose you face a two-armed adversarial bandit with losses $y_{t1} = \mathbb{I}\{t \leq n/2\}$ and $y_{t2} = \mathbb{I}\{t > n/2\}$. If we run Exp3 on this problem, then Theorem 11.2 guarantees that

$$R_n = \mathbb{E} \left[\sum_{t=1}^n y_{tA_t} \right] - \min_{i \in \{1,2\}} \sum_{t=1}^n y_{ti} \leq \sqrt{2nK \log(K)}.$$

Since $\min_{i \in \{1,2\}} \sum_{t=1}^n y_{ti} = n/2$, by rearranging we see that

$$\mathbb{E} \left[\sum_{t=1}^n y_{tA_t} \right] \leq \frac{n}{2} + \sqrt{2nK \log(K)}.$$

To put this in perspective, a policy that plays each arm with probability half in every round would have $\mathbb{E}[\sum_{t=1}^n y_{tA_t}] = n/2$. In other words, the regret guarantee is practically meaningless.

What should we expect for this problem? The sequence of losses is so regular that we might hope our policy will mostly play the second arm in the first $n/2$ rounds and then switch to playing mostly the first arm in the second $n/2$ rounds. Then the cumulative loss would be close to zero and the regret would be negative. Rather than trying to prove a negative regret, let us redefine the regret

to strengthen the competitor class. Let $\Gamma_m \subset [K]^n$ be the set of action sequences with at most $m - 1$ changes.

$$\Gamma_m = \left\{ (a_t) \in [K]^n : \sum_{t=1}^{n-1} \mathbb{I}\{a_t \neq a_{t+1}\} \leq m - 1 \right\}$$

Then define the **nonstationary regret** with $m - 1$ change points by

$$R_{n,m} = \mathbb{E} \left[\sum_{t=1}^n y_{tA_t} \right] - \min_{a \in \Gamma_m} \mathbb{E} \left[\sum_{t=1}^n y_{ta_t} \right].$$

The nonstationary regret is sometimes called the **tracking regret** because a learner that makes it small must ‘track’ the best arm as it changes. Notice that $R_{n,1}$ coincides with the usual definition of the regret. Furthermore, on the sequence described at the beginning of the section we see that

$$R_{n,2} = \mathbb{E} \left[\sum_{t=1}^n y_{tA_t} \right],$$

which means a policy can only enjoy sublinear nonstationary regret if it detects the change point quickly. The obvious question is whether or not such a policy exists and how its regret depends on m .

Exp4 for nonstationary bandits

One idea is to use the Exp4 policy from Chapter 18 with a large set of experts, one for each $a \in \Gamma_m$. Theorem 18.1 shows that Exp4 with these experts suffers regret of at most

$$R_{n,m} \leq \sqrt{2nK \log |\Gamma_m|}.$$

Naively bounding $\log |\Gamma_m|$ and ignoring constant factors shows that

$$R_{n,m} = O \left(\sqrt{nmK \log \left(\frac{Kn}{m} \right)} \right).$$

To see that you cannot do much better than this, imagine playing m adversarial bandits sequentially, each with horizon n/m . No matter what policy you propose, there exist choices of bandits such that the expected regret suffered against each bandit is at least $\Omega(\sqrt{nK/m})$. And after summing over the m instances we see that the worst case regret is at least

$$R_{n,m} = \Omega \left(\sqrt{nmK} \right),$$

which matches the upper bound except for logarithmic factors. Notice how this lower bound applies to policies that know the location of the changes, so it is not true that things are significantly harder in the absence of this knowledge. There is one big caveat with all these calculations. The running time of a naive implementation of Exp4 is linear in the number of experts, which even for modestly sized m is very large indeed.

Online Stochastic Mirror Descent

The computational issues faced by Exp4 are most easily overcome using the tools from online convex optimization developed in Chapter 28. The idea is to use online stochastic mirror descent and the negentropy potential. Without further modification this would be Exp3, which you will show does not work for nonstationary bandits (Exercise 31.2). The trick is to restrict the action set to the clipped simplex $\mathcal{A} = \mathcal{P}_{K-1} \cap [\alpha, 1]^K$ where $\alpha \in [0, 1/K]$ is a constant to be tuned subsequently. The clipping ensures the algorithm does not commit too hard to any single arm, which prevents it from discovering the change points. Let $F : [0, \infty)^K \rightarrow \mathbb{R}$ be the unnormalized negentropy potential and $P_1 \in \mathcal{A}$ be the uniform probability vector. In each round t the learner samples $A_t \sim P_t$ and updates its sampling distribution by

$$P_{t+1} = \operatorname{argmin}_{p \in \mathcal{A}} \eta \langle p, \hat{Y}_t \rangle + D_F(p, P_t), \quad (31.1)$$

where $\eta > 0$ is the learning rate and $\hat{Y}_{ti} = \mathbb{I}\{A_t = i\} y_{ti}/P_{ti}$ is the importance-weighted estimator. The optimization problem in Eq. (31.1) can be computed efficiently using the two-step process:

$$\begin{aligned} \tilde{P}_{t+1} &= \operatorname{argmin}_{p \in [0, \infty)^K} \eta \langle p, \hat{Y}_t \rangle + D_F(p, P_t) \\ P_{t+1} &= \operatorname{argmin}_{p \in \mathcal{A}} D_F(p, \tilde{P}_{t+1}). \end{aligned}$$

The first of these subproblems can be evaluated analytically, yielding $\tilde{P}_{t+1,i} = P_{ti} \exp(-\eta \hat{Y}_{ti})$. The second can be solved efficiently using the result in Exercise 26.8.

THEOREM 31.1 *The regret of the policy sampling $A_t \sim P_t$ with P_t defined in Eq. (31.1) is bounded by*

$$R_{n,m} \leq \alpha n(K-1) + \frac{m \log(1/\alpha)}{\eta} + \frac{\eta n K}{2}.$$

Proof Let $a^* \in \operatorname{argmin}_{a \in \Gamma_m} \sum_{t=1}^n y_{ta}$ be an optimal sequence of actions in hindsight constrained to Γ_m . Then let $1 = t_1 < t_2 < \dots < t_m < t_{m+1} = n$ so that a_t^* is constant on each interval $\{t_i, \dots, t_{i+1}-1\}$. We abuse notation by writing $a_i^* = a_{t_i}^*$. Then the regret decomposes into

$$\begin{aligned} R_{n,m} &= \mathbb{E} \left[\sum_{t=1}^n (y_{tA_t} - y_{ta_t^*}) \right] = \mathbb{E} \left[\sum_{i=1}^m \sum_{t=t_i}^{t_{i+1}-1} (y_{tA_t} - y_{ta_i^*}) \right] \\ &= \sum_{i=1}^m \mathbb{E} \left[\mathbb{E} \left[\sum_{t=t_i}^{t_{i+1}-1} (y_{tA_t} - y_{ta_i^*}) \mid P_{t_i} \right] \right]. \end{aligned}$$

The next step is to apply Theorem 28.1 and the solution to Exercise 28.5 to

bound the inner expectation.

$$\begin{aligned} \mathbb{E} \left[\sum_{t=t_i}^{t_{i+1}-1} (y_{tA_t} - y_{ta_i^*}) \mid P_{t_i} \right] &= \mathbb{E} \left[\sum_{t=t_i}^{t_{i+1}-1} \langle P_t - e_{a_i^*}, y_t \rangle \mid P_{t_i} \right] \\ &\leq \alpha(t_{i+1} - t_i)(K-1) + \mathbb{E} \left[\max_{p \in \mathcal{A}} \sum_{t=t_i}^{t_{i+1}-1} \langle P_t - p, y_t \rangle \mid P_{t_i} \right] \\ &\leq \alpha(t_{i+1} - t_i)(K-1) + \mathbb{E} \left[\max_{p \in \mathcal{A}} \frac{D(p, P_{t_i})}{\eta} + \frac{\eta K(t_{i+1} - t_i)}{2} \mid P_{t_i} \right]. \end{aligned}$$

By assumption $P_{t_i} \in \mathcal{A}$ and so $P_{t,j} \geq \alpha$ for all j and $D(p, P_{t_i}) \leq \log(1/\alpha)$. Combining this observation with the previous two displays shows that

$$R_{n,m} \leq n\alpha(K-1) + \frac{m \log(1/\alpha)}{\eta} + \frac{\eta n K}{2}. \quad \square$$

The learning rate and clipping parameters are approximately optimized by

$$\eta = \sqrt{2m \log(1/\alpha)/(nK)} \quad \text{and} \quad \alpha = \sqrt{m/(nK)},$$

which leads to a regret of $R_{n,m} \leq \sqrt{mnK \log(nK/m)} + \sqrt{mnK}$. In typical applications the value of m is not known. In this case one can choose $\eta = \sqrt{\log(1/\alpha)/nK}$ and $\alpha = \sqrt{1/nK}$ and the regret increases by a factor of $O(\sqrt{m})$.

31.2 Stochastic bandits

We saw in Part II that by making a statistical assumption on the rewards it was possible to design policies with logarithmic regret. This is the big advantage of making assumptions – you get stronger results. The nonstationarity makes the modelling problem less trivial. To keep things simple we will assume the rewards are Gaussian and that for each arm i there is a function $\mu_i : [n] \rightarrow \mathbb{R}$ and the reward is

$$X_t = \mu_{A_t}(t) + \eta_t,$$

where (η_t) is a sequence of independent standard Gaussian random variables. The optimal arm in round t has mean $\mu^*(t) = \max_{i \in [K]} \mu_i(t)$ and the regret is

$$R_n(\mu) = \sum_{t=1}^n \mu^*(t) - \mathbb{E} \left[\sum_{t=1}^n \mu_{A_t}(t) \right].$$

The amount of nonstationarity is modelled by placing restrictions on the functions μ_i . To be consistent with the previous section we assume the mean vector changes at most $m-1$ times, which amounts to saying that

$$\sum_{t=1}^{n-1} \max_{i \in [K]} \mathbb{I}\{\mu_i(t) \neq \mu_i(t+1)\} \leq m-1.$$

Suppose the locations of the change points were known, then running a new copy of UCB on each interval would lead to a bound of

$$R_n(\mu) = O\left(\frac{mK}{\Delta_{\min}} \log\left(\frac{n}{m}\right)\right),$$

where Δ_{\min} is the smallest suboptimality gap over all m blocks. In the last section we saw the bound achieved by an omniscient policy that knows when the changes occur can be achieved by a policy that does not. Unfortunately this is not true here.

THEOREM 31.2 *Let $K = 2$ and suppose that $\mu_i(t) = \mu_i$ is constant for both arms and $\Delta = \mu_2 - \mu_1 > 0$. Then for all sufficiently large n there exists a nonstationary bandit μ' with two change points such that $R_n(\mu') \geq cn/R_n(\mu)$, where $c > 0$ is a universal constant.*

The theorem shows that if a policy enjoys $R_n(\mu) = o(n^{1/2})$ for any nontrivial (stationary) bandit, then its minimax regret is at least $\omega(n^{1/2})$ on some nonstationary bandit. In particular, if $R_n(\mu) = O(\log(n))$, then the minimax regret is at least $\Omega(n/\log(n))$. This immediately dashes our hopes for a policy that is much better than Exp4 in a stochastic setting. There are algorithms designed for nonstationary bandits in the stochastic setting with abrupt change points as described above. Those that come with theoretical guarantees are based on forgetting or discounting data so that decisions of the algorithm depend almost entirely on recent data. In the notes we discuss these approaches along with alternative models for nonstationarity.

Proof of Theorem 31.2 Let $(S_k)_{k=1}^L$ be a partition of $[n]$ to be specified later. Let \mathbb{P} and $\mathbb{E}[\cdot]$ denote the probabilities and expectations with respect to the bandit determined by μ and \mathbb{P}' with respect to alternative nonstationary bandit μ' to be defined shortly. By the pigeonhole principle there exists a $k \in [L]$ such that

$$\mathbb{E}\left[\sum_{t \in S_k} \mathbb{I}\{A_t = 2\}\right] \leq \frac{\mathbb{E}[T_2(n)]}{L}.$$

Define an alternative nonstationary bandit with $\mu'(t) = \mu$ except for $t \in S_k$ when we let $\mu'_2(t) = \mu_2 + \varepsilon$ where $\varepsilon = \sqrt{2L/\mathbb{E}[T_2(n)]}$. Then by Lemma 15.1 and Theorem 14.2,

$$\begin{aligned} \mathbb{P}\left(\sum_{t \in S_k} \mathbb{I}\{A_t = 2\} \geq \frac{|S_k|}{2}\right) + \mathbb{P}'\left(\sum_{t \in S_k} \mathbb{I}\{A_t = 2\} < \frac{|S_k|}{2}\right) &\geq \frac{1}{2} \exp(-D(\mathbb{P}, \mathbb{P}')) \\ &\geq \frac{1}{2} \exp\left(-\frac{\mathbb{E}[T_2(n)]\varepsilon^2}{2L}\right) \geq \frac{1}{2e}, \end{aligned}$$

By Markov's inequality,

$$\mathbb{P}\left(\sum_{t \in S_k} \mathbb{I}\{A_t = 2\} \geq \frac{|S_k|}{2}\right) \leq \frac{2}{|S_k|} \mathbb{E}\left[\sum_{t \in S_k} \mathbb{I}\{A_t = 2\}\right] \leq \frac{2\mathbb{E}[T_2(n)]}{L|S_k|} \leq \frac{1}{\Delta^2|S_k|},$$

where the last inequality follows by choosing $L = \lceil 2\Delta^2\mathbb{E}[T_2(n)] \rceil$, which also ensures that $\varepsilon - \Delta \geq \varepsilon/2$. Therefore

$$R_n(\mu') \geq \left(\frac{1}{2e} - \frac{1}{\Delta^2|S_k|}\right) \frac{\varepsilon|S_k|}{4} = \left(\frac{1}{2e} - \frac{1}{\Delta^2|S_k|}\right) \frac{|S_k|\Delta}{2}.$$

If (S_k) is chosen as a uniform partition so that $|S_k| \geq \lfloor n/L \rfloor$, then there exists a universal constant $c > 0$ such that for sufficiently large n , $R_n(\mu') \geq cn/R_n(\mu)$. \square

31.3 Notes

1 The negative results for stochastic nonstationary bandits do not mean that trying to improve on the adversarial bandit algorithms is completely hopeless. First of all, the adversarial bandit algorithms are not well suited for exploiting distributional assumptions on the noise, which makes things irritating when the losses/rewards are Gaussian (which are unbounded) or Bernoulli (which have small variance near the boundaries). There have been several algorithms designed specifically for stochastic nonstationary bandits. When the reward distributions are permitted to change abruptly as in the last section, then the two main algorithms are based on the idea of ‘forgetting’ rewards observed in the distant past. One way to do this is with **discounting**. Let $\gamma \in (0, 1)$ be the **discount factor** and define

$$\hat{\mu}_i^\gamma(t) = \sum_{s=1}^t \gamma^{t-s} \mathbb{I}\{A_s = i\} X_s \quad T_i^\gamma(t) = \sum_{s=1}^t \gamma^{t-s} \mathbb{I}\{A_s = i\}.$$

Then for appropriately tuned constant α the Discounted UCB policy chooses each arm once and subsequently

$$A_t = \operatorname{argmax}_{i \in [K]} \left(\hat{\mu}_i^\gamma(t-1) + \sqrt{\frac{\alpha}{T_i^\gamma(t-1)} \log \left(\sum_{i=1}^K T_i^\gamma(t-1) \right)} \right).$$

The idea is to ‘discount’ rewards that occurred far in the past, which makes the algorithm most influenced by recent events. A similar algorithm called Sliding-Window UCB uses a similar approach, but rather than discounting past rewards with a geometric discount function it simply discards them altogether. Let $\tau \in \mathbb{N}^+$ be a constant and define

$$\hat{\mu}_i^\tau(t) = \sum_{s=t-\tau+1}^t \mathbb{I}\{A_s = i\} X_s \quad T_i^\tau(t) = \sum_{s=t-\tau+1}^t \mathbb{I}\{A_s = i\}.$$

Then the Sliding-Window UCB chooses

$$A_t = \operatorname{argmax}_{i \in [K]} \left(\hat{\mu}_i^\tau(t-1) + \sqrt{\frac{\alpha}{T_i^\tau(t-1)} \log(t \wedge \tau)} \right).$$

It is known that if γ or τ are tuned appropriately, then for Discounted UCB the regret may be bounded by $O(\sqrt{nm} \log(n))$ and for Sliding-Window UCB by $O(\sqrt{nm \log(n)})$. Neither bound improves on what is available using Exp4, but there is some empirical evidence to support the use of these algorithms when the stochastic assumption holds.

- 2 An alternative way to model nonstationary stochastic bandits is to assume the mean payoffs of the arms are slowly drifting. One way to do this is to assume that $\mu_i(t)$ follows a reflected Brownian motion in some interval. It is not hard to see that the regret is necessarily linear in this case because the best arm can change in any round with nonzero nondecreasing probability. The objective in this case is to understand the magnitude of the linear regret in terms of the size of the interval or volatility of the Brownian motion.
- 3 Yet another idea is to allow the means to change in an arbitrary way, but restrict the amount of variation. Let $\mu_t = (\mu_1(t), \dots, \mu_K(t))$ and

$$V_n = \sum_{t=1}^{n-1} \|\mu_t - \mu_{t+1}\|_\infty$$

be the cumulative change in mean rewards measured in terms of the supremum norm. Then for each $V \in [1/K, n/K]$ there exists a policy such that for all bandits with $V_n \leq V$ it holds that

$$R_n \leq C(VK \log(K))^{1/3} T^{2/3}.$$

And furthermore, this bound is nearly tight in a minimax sense except for logarithmic terms [Besbes et al., 2014].

31.4 Bibliographic remarks

Nonstationary bandits have quite a long history. The celebrated Gittins index is based on a model where each arm is associated with a Markov chain that evolves when played and the reward depends on the state [Gittins, 1979, Gittins et al., 2011]. The classical approaches address this problem in the Bayesian framework and the objective is primarily to design efficient algorithms rather than understanding the frequentist regret. Note that the state is observed after each action. Even more related is the **restless bandit**, which is the same as Gittin's setup except the Markov chain for every action evolves in every round. The problem is made challenging because the learner still only observes the state and reward for the action they chose. Restless bandits were introduced by Whittle [1988] in the Bayesian framework and unfortunately there are more negative results than positive ones. There has been some interesting frequentist analysis,

but the challenging nature of the problem makes it difficult to design efficient algorithms with meaningful regret guarantees [Ortner et al., 2012]. Certainly there is potential for more work in this area. The ideas in Section 31.1 are mostly generalizations of algorithms designed for the full information setting, notably the Fixed Share algorithm Herbster and Warmuth [1998]. The first algorithm designed for the adversarial nonstationary bandit is Exp3.S by Auer et al. [2002b]. This algorithm can be interpreted as an efficient version of Exp4 with a carefully chosen initialization such that the exponential computation over all experts collapses into a simple expression. We do not know of a clean source for this interpretation, but see the analysis of Fixed Share in the book by Cesa-Bianchi and Lugosi [2006]. The Exp3.P policy was originally developed in order to prove high probability bounds for finite-armed adversarial bandits [Auer et al., 2002b], but Audibert and Bubeck [2010b] proved that with appropriate tuning it also enjoys the same bounds as Exp3.S. Presumably this also holds for Exp3-IX. Mirror descent has been used to prove tracking bounds in the full information setting by Herbster and Warmuth [2001]. A more recent reference is by György and Szepesvári [2016], which makes the justification for clipping explicit. The lower bound for stochastic nonstationary bandits is by Garivier and Moulines [2011], though our proof differs in minor ways. We mentioned that there is a line of work on stochastic nonstationary bandits where the rewards are slowly drifting. The approach based on Brownian motion is due to Slivkins and Upfal [2008] while the variant described in Note 3 is by Besbes et al. [2014]. The idea of discounted UCB was introduced without analysis by Kocsis and Szepesvári [2006]. The analysis of this algorithm and also the sliding window algorithm is by Garivier and Moulines [2011].

31.5 Exercises

31.1 Let $n, m, K \in \mathbb{N}^+$ be such that $n \geq mK$. Prove that for any policy π there exists an adversarial bandit (y_{ti}) such that

$$R_{n,m} \geq c\sqrt{nmK},$$

where $c > 0$ is a universal constant.

31.2 Prove for all sufficiently large n that Exp3 from Chapter 11 has $R_{n,2} \geq cn$ for some universal constant $c > 0$.

31.3 Let $K = 2$ and $n = 1000$ and define adversarial bandit in terms of losses with $y_{t1} = \mathbb{I}\{t < n/2\}$ and $y_{t2} = \mathbb{I}\{t \geq n/2\}$. Plot the expected regret of Exp3, Exp3-IX and the variant of online stochastic mirror descent proposed in this chapter. Experiment with a number of learning rates for each algorithm.

32 Ranking

Ranking is the process of producing an ordered shortlist of K items from a larger collection of L items. These tasks come in several flavors. Sometimes the user supplies a query and the system responds with a shortlist of items. In other applications the shortlist is produced without an explicit query. For example, a streaming service might provide a list of recommended movies when you sign in. Our focus here is on the second type of problem.

We examine a sequential version of the ranking problem where the learner selects a ranking, receives feedback about its quality and repeats the process over n rounds. The feedback will be in the form of ‘clicks’ from the user, which comes from the view that ranking is a common application in on-line recommendation systems and the user selects the items they like by clicking on them. The objective of the learner is to maximize the expected number of clicks.

A **permutation** on $[L]$ is an invertible function $\sigma : [L] \rightarrow [L]$. Let \mathcal{A} be the set of all permutations on $[L]$. In each round t the learner chooses an action $A_t \in \mathcal{A}$, which should be interpreted as meaning the learner places item $A_t(k)$ in the k th position. Equivalently, $A_t^{-1}(i)$ is the position of the i th item. Since the shortlist has length K the order of $A_t(K+1), \dots, A_t(L)$ is not important and is included only for notational convenience. After choosing their action, the learner observes $C_{ti} \in \{0, 1\}$ for each $i \in [L]$ where $C_{ti} = 1$ if the user clicked on the i th item in the collection. Note that the user may click on multiple items. We will assume a stochastic model where the probability that the user clicks on position k in round t only depends on A_t and is given by $v(A_t, k)$ with $v : \mathcal{A} \times [L] \rightarrow [0, 1]$ an unknown function. The regret over n rounds is

$$R_n = n \max_{a \in \mathcal{A}} \sum_{k=1}^L v(a, k) - \mathbb{E} \left[\sum_{t=1}^n \sum_{i=1}^L C_{ti} \right].$$

A naive way to minimize the regret would be to create a finite-armed bandit where each arm corresponds to a ranking of the items and then apply your favourite algorithm from Part II. The problem is that these algorithms treat the arms as independent and cannot exploit any structure in the ranking. This is almost always unacceptable because the number of ways to rank K items from a collection of size L is $L!/(L-K)!$. Ranking illustrates of one of the most fundamental dilemmas in machine learning: choosing a model. A rich model leads to low misspecification error, but takes longer to fit while a course model

can suffer from large misspecification error. In the context of ranking a model corresponds to assumptions on the function v .

32.1 Click models

The only way to avoid the curse of dimensionality is to make assumptions. A natural way to do this for ranking is to assume that the probability of clicking on an item depends on (a) the underlying quality of that item and (b) the location of that item in the chosen ranking. A formal definition of how this is done is called a **click model**. Deciding which model to use depends on the particulars of the problem at hand, such as how the list is presented to the user and whether or not clicking on an item diverts them to a different page. This issue has been studied by the data retrieval community and there is now a large literature devoted to the pros and cons of different choices. We limit ourselves to describing the popular choices and give pointers to the literature at the end of the chapter.

Document-based model

The **document-based model** is one of the simplest click models, which assumes the probability of clicking on a shortlisted item is equal to its **attractiveness**. Formally, for each item $i \in [L]$ let $\alpha(i) \in [0, 1]$ be the attractiveness of item i . The document-based model assumes that

$$v(a, k) = \alpha(a(k))\mathbb{I}\{k \leq K\} .$$

The unknown quantity in this model is the attractiveness function, which has just L parameters.

Position-based model

The document-based model might occasionally be justified, but in most cases the position of an item in the ranking also affects the likelihood of a click. A natural extension that accounts for this behavior is called the **position-based model**, which assumes that

$$v(a, k) = \alpha(a(k))\chi(k) ,$$

where $\chi : [L] \rightarrow [0, 1]$ is a function that measures the quality of position k . Since the user cannot click on items that are not shown we assume that $\chi(k) = 0$ for $k > K$. This model is richer than the document-based model, which is recovered by choosing $\chi(k) = \mathbb{I}\{k \leq K\}$. The number of parameters in the position-based models is $K + L$.

Cascade model

The position-based model is not suitable for applications where clicking on an item takes the user to a different page. In the **cascade model** it is assumed that

the learner scans the shortlisted items in order and only clicks on the first item they find attractive. Define $\chi : \mathcal{A} \times [L] \rightarrow [0, 1]$ by

$$\chi(a, k) = \begin{cases} 1 & \text{if } k = 1 \\ 0 & \text{if } k > K \\ \prod_{\ell=1}^{k-1} (1 - \alpha(a(\ell))) & \text{otherwise,} \end{cases}$$

which is the probability that the user has not clicked on the first $k - 1$ items. Then the cascade model assumes that

$$v(a, k) = \alpha(a(k))\chi(a, k). \quad (32.1)$$

The first term in the factorization is the attractiveness function, which measures the probability that the user is attracted to the i th item. The second term can be interpreted as the probability that the user examines that item. This interpretation is also valid in the position-based model. It is important to emphasize that $v(a, k)$ is the probability of clicking on the k th position when taking action $a \in \mathcal{A}$. This does not mean that C_{t1}, \dots, C_{tL} are independent. So far the assumptions only restrict the marginal distribution of each C_{ti} , which for most of this chapter is all we require. Nevertheless, in the cascade model it would be standard to assume that $C_{tA_t(k)} = 0$ if there exists an $\ell < k$ such that $C_{tA_t(\ell)} = 1$ and otherwise

$$\mathbb{P}(C_{tA_t(k)} = 1 \mid A_t, C_{tA_t(1)} = 0, \dots, C_{tA_t(k-1)} = 0) = \mathbb{I}\{k \leq K\} \alpha(A_t(k)).$$

Like the document-based model, the cascade model has L parameters.

Generic model

We now introduce a model that generalizes the last three. Previous models essentially assumed that the probability of a click factorizes into an attractiveness probability and an examination probability. We deviate from this norm by making assumptions directly on the function v . Given $\alpha : [L] \rightarrow [0, 1]$, an action a is called α -optimal if the shortlisted items are the K most attractive sorted by attractiveness: $\alpha(a(k)) = \max_{k' > k} \alpha(a(k'))$ for all $k \in [K]$.

ASSUMPTION 32.1 There exists an attractiveness function $\alpha : [L] \rightarrow [0, 1]$ such that the following four conditions are satisfied. Let $a \in \mathcal{A}$ and $i, j, k \in [L]$ be such that $\alpha(i) \geq \alpha(j)$ and let σ be the permutation that exchanges i and j .

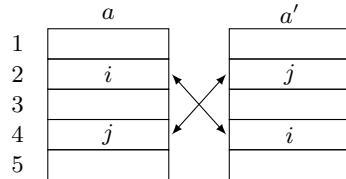
- (a) $v(a, k) = 0$ for all $k > K$.
- (b) $\sum_{k=1}^K v(a^*, k) = \max_{a \in \mathcal{A}} \sum_{k=1}^K v(a, k)$ for all α -optimal actions a^* .
- (c) For all i and j with $\alpha(i) \geq \alpha(j)$

$$v(a, a^{-1}(i)) \geq \frac{\alpha(i)}{\alpha(j)} v(\sigma \circ a, a^{-1}(i)),$$

where σ is the permutation on $[L]$ that exchanges i and j .

- (d) If a is an action such that $\alpha(a(k)) = \alpha(a^*(k))$ for some α -optimal action a^* , then $v(a, k) \geq v(a^*, k)$.

These assumptions may appear quite mysterious. At some level they are chosen to make the proof go through, while simultaneously generalizing the document-based, position-based and cascade models (32.1). The choices not entirely without basis or intuition, however. Part (a) asserts that the user does not click on items that are not placed in the shortlist. Part (b) says that α -optimal actions maximize the expected number of clicks. Note that there are multiple optimal rankings if α is not injective. Part (c) is a little more restrictive and is illustrated in the figure. The probability of clicking on the second position is larger in the ranking on the left than the right by a factor of at least $\alpha(i)/\alpha(j)$. On the other hand, the probability of clicking on the fourth position is larger in the ranking on the right. One way to justify this is to assume that $v(a, k) = \alpha(a(k))\chi(a, k)$ where $\chi(a, k)$ is viewed as the probability that the user examines position k . It seems reasonable to assume that the probability the user examines position k should only depend on the first $k - 1$ items. Hence $v(a, 2) = \alpha(i)\chi(a, 2) = \alpha(i)\chi(a', 2) = \alpha(i)/\alpha(j)v(a', 2)$. In order to make the argument for the fourth position we need to assume that placing less attractive items in the early slots increases the probability that the user examines later positions (searching for a good result). This is true for the position-based and cascade models, but is perhaps the most easily criticised assumption. Part (d) says that the probability that a user clicks on a position with a correctly placed item is at least as large as the probability that the user clicks on that position in an optimal ranking. The justification is that the items $a(1), \dots, a(k - 1)$ cannot be more attractive than $a^*(1), \dots, a^*(k - 1)$, which should increase the likelihood that the user makes it the k th position.



The generic model has many parameters, but we will see that the learner does not need to learn all of them in order to suffer small regret. The advantage of this model relative to the previous ones is that it offers more flexibility and yet it is not so flexible that learning is impossible.

32.2 Policy

We now explain the policy for learning to rank when v is unknown, but satisfies Assumption 32.1. After the description is an illustration that may prove helpful.

Step 0: Initialization

The policy takes as input a confidence parameter $\delta \in (0, 1)$ and L and K . The policy maintains a binary relation $G_t \subseteq [L] \times [L]$. In the first round $t = 1$ the relation is empty: $G_1 = \emptyset$. You should think of G_t as maintaining pairs (i, j) for which the policy has proven with high probability that $\alpha(i) < \alpha(j)$. Ideally, $G_t \subseteq G = \{(i, j) \in [L] \times [L] : \alpha(i) \leq \alpha(j)\}$.

Step 1: Defining a partition

In each round t the learner computes a partition of the actions based on a topological sort according to relation G_t . Given $A \subset [L]$ define $\min_{G_t}(A)$ to be the set of minimum elements of A according to relation G_t .

$$\min_{G_t}(A) = \{i \in A : (i, j) \notin G_t \text{ for all } j \in G_t\}.$$

Then let $\mathcal{P}_{t1}, \mathcal{P}_{t2}, \dots$ be the partition of $[L]$ defined inductively by

$$\mathcal{P}_{td} = \min_{G_t} \left([L] \setminus \bigcup_{c=1}^{d-1} \mathcal{P}_{tc} \right).$$

Finally, let $M_t = \max\{d : \mathcal{P}_{td} \neq \emptyset\}$. The reader should check that if G_t does not have cycles, then M_t is well defined and finite and that $\mathcal{P}_{t1}, \dots, \mathcal{P}_{tM_t}$ is indeed a partition of $[L]$ (Exercise 32.5). The event that G_t contains cycles is a failure event. In order for the policy to be well defined we assume it chooses some arbitrary fixed action in this case.

Step 2: Choosing an action

Define $\mathcal{I}_{t1}, \dots, \mathcal{I}_{tM_t}$ be the partition of $[L]$ defined inductively by

$$\mathcal{I}_{td} = [\cup_{c \leq d} \mathcal{P}_{tc}] \setminus [\cup_{c < d} \mathcal{P}_{tc}].$$

Next let $\Sigma_t \subseteq \mathcal{A}$ be the set of actions σ such that $\sigma(\mathcal{I}_{td}) = \mathcal{P}_{td}$ for all $d \in [M_t]$. The algorithm chooses A_t uniformly at random from Σ_t . Intuitively the policy first shuffles the items in \mathcal{P}_{t1} and uses these as the first $|\mathcal{P}_{t1}|$ entries in the ranking. Then \mathcal{P}_{t2} is shuffled and the items are appended to the ranking. This process is repeated until the ranking is complete. For an item $i \in [L]$, we denote by D_{ti} the unique index d such that $i \in \mathcal{P}_{td}$.

Step 3: Updating the relation

For any pair of items $i, j \in [L]$ define $S_{tij} = \sum_{s=1}^t U_{sij}$ and $N_{tij} = \sum_{s=1}^t |U_{sij}|$ where

$$U_{tij} = \mathbb{I}\{D_{ti} = D_{tj}\} (C_{ti} - C_{tj}).$$

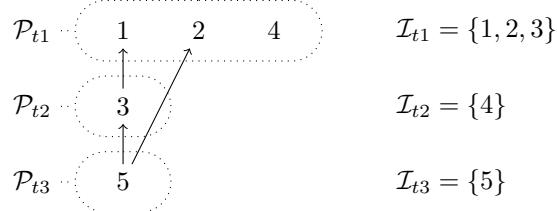
All this means is that S_{tij} tracks the differences between the number of clicks of items i and j over rounds when they share a partition. As a final step, the relation G_{t+1} is given by

$$G_{t+1} = G_t \cup \left\{ (j, i) : S_{tij} \geq \sqrt{2N_{tij} \log \left(\frac{c\sqrt{N_{tij}}}{\delta} \right)} \right\},$$

where $c \approx 3.43$ is the universal constant given in Exercise 20.7. In the analysis we will show that if $\alpha(i) \geq \alpha(j)$, then with high probability S_{tji} is never large enough for G_{t+1} to include (i, j) . In this sense, with high probability G_t is consistent with the order on $[L]$ induced by sorting in decreasing order with respect to $\alpha(\cdot)$. Note that G_t is generally not a partial order because it is not transitive.

Illustration

Suppose $L = 5$ and $K = 4$ and in round t the relation is $G_t = \{(3, 1), (5, 2), (5, 3)\}$, which is represented in the graph below where an arrow from j to i indicates that $(j, i) \in G_t$.



This means that in round t the first three positions in the ranking will contain items from $\mathcal{P}_{t1} = \{1, 2, 4\}$, but with random order. The fourth position will be item 3 and item 5 is not shown to the user.



Part (a) of Assumption 32.1 means that items in position $k > K$ are never clicked. As a consequence, the algorithm never needs to actually compute the partitions \mathcal{P}_{td} for which $\min \mathcal{I}_{td} > K$ because items in these partitions are never shortlisted.

32.3 Regret analysis

THEOREM 32.1 *Let function v satisfy Assumption 32.1 and assume that $\alpha(1) > \alpha(2) > \dots > \alpha(L)$. Let $\Delta_{ij} = \alpha(i) - \alpha(j)$ and $\delta \in (0, 1)$. Then the regret of TOPRANK is bounded from above as*

$$R_n \leq \delta n K L^2 + \sum_{j=1}^L \sum_{i=1}^{\min\{K, j-1\}} \left(1 + \frac{6(\alpha(i) + \alpha(j)) \log\left(\frac{c\sqrt{n}}{\delta}\right)}{\Delta_{ij}} \right).$$

$$\text{Furthermore, } R_n \leq \delta n K L^2 + KL + \sqrt{4K^3 L n \log\left(\frac{c\sqrt{n}}{\delta}\right)}.$$

By choosing $\delta = n^{-1}$ the theorem shows that the expected regret is at most

$$R_n = O\left(\sum_{j=1}^L \sum_{i=1}^{\min\{K, j-1\}} \frac{\alpha(i) \log(n)}{\Delta_{ij}}\right) \quad \text{and} \quad R_n = O\left(\sqrt{K^3 L n \log(n)}\right).$$

The algorithm does not make use of any assumed ordering on $\alpha(\cdot)$, so the assumption is only used to allow for a simple expression for the regret. The core idea of the proof is to show that (a) if the algorithm is suffering regret as a consequence of misplacing an item, then it is gaining information about the relation of the items so that G_t will gain elements and (b) once G_t is sufficiently

rich the algorithm is playing optimally. Let $\mathcal{F}_t = \sigma(A_1, C_1, \dots, A_t, C_t)$ and $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \mid \mathcal{F}_t)$ and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. For each $t \in [n]$ let F_t to be the failure event that there exists $i \neq j \in [L]$ and $s < t$ such that $N_{sij} > 0$ and

$$\left| S_{sij} - \sum_{u=1}^s \mathbb{E}_{u-1}[U_{uji} \mid U_{uji} \neq 0] |U_{uji}| \right| \geq \sqrt{2N_{sij} \log(c\sqrt{N_{sij}}/\delta)}.$$

LEMMA 32.1 *Let i and j satisfy $\alpha(i) \geq \alpha(j)$ and $d \geq 1$. On the event that $i, j \in \mathcal{P}_{sd}$ and $d \in [M_s]$ and $U_{sij} \neq 0$, the following hold almost surely:*

- (a) $\mathbb{E}_{s-1}[U_{sij} \mid U_{sij} \neq 0] \geq \frac{\Delta_{ij}}{\alpha(i) + \alpha(j)}$.
- (b) $\mathbb{E}_{s-1}[U_{sji} \mid U_{sji} \neq 0] \leq 0$.

Proof For the remainder of the proof we focus on the event that $i, j \in \mathcal{P}_{sd}$ and $d \in [M_s]$ and $U_{sij} \neq 0$. We also discard the measure zero subset of this event where $\mathbb{P}_{s-1}(U_{sij} \neq 0) = 0$. From now on we omit the ‘almost surely’ qualification on conditional expectations. Under these circumstances the definition of conditional expectation shows that

$$\begin{aligned} \mathbb{E}_{s-1}[U_{sij} \mid U_{sij} \neq 0] &= \frac{\mathbb{P}_{s-1}(C_{si} = 1, C_{sj} = 0) - \mathbb{P}_{s-1}(C_{si} = 0, C_{sj} = 1)}{\mathbb{P}_{s-1}(C_{si} \neq C_{sj})} \\ &= \frac{\mathbb{P}_{s-1}(C_{si} = 1) - \mathbb{P}_{s-1}(C_{sj} = 1)}{\mathbb{P}_{s-1}(C_{si} \neq C_{sj})} \\ &\geq \frac{\mathbb{P}_{s-1}(C_{si} = 1) - \mathbb{P}_{s-1}(C_{sj} = 1)}{\mathbb{P}_{s-1}(C_{si} = 1) + \mathbb{P}_{s-1}(C_{sj} = 1)} \\ &= \frac{\mathbb{E}_{s-1}[v(A_s, A_s^{-1}(i)) - v(A_s, A_s^{-1}(j))]}{\mathbb{E}_{s-1}[v(A_s, A_s^{-1}(i)) + v(A_s, A_s^{-1}(j))]}, \end{aligned} \quad (32.2)$$

where in the second equality we added and subtracted $\mathbb{P}_{s-1}(C_{si} = 1, C_{sj} = 1)$. By the design of TOPRANK, the items in \mathcal{P}_{td} are placed into slots \mathcal{I}_{td} uniformly at random. Let σ be the permutation that exchanges the positions of items i and j . Then using Part Item (c) of Assumption 32.1,

$$\begin{aligned} \mathbb{E}_{s-1}[v(A_s, A_s^{-1}(i))] &= \sum_{a \in \mathcal{A}} \mathbb{P}_{s-1}(A_s = a) v(a, a^{-1}(i)) \\ &\geq \frac{\alpha(i)}{\alpha(j)} \sum_{a \in \mathcal{A}} \mathbb{P}_{s-1}(A_s = a) v(\sigma \circ a, a^{-1}(i)) \\ &= \frac{\alpha(i)}{\alpha(j)} \sum_{a \in \mathcal{A}} \mathbb{P}_{s-1}(A_s = \sigma \circ a) v(\sigma \circ a, (\sigma \circ a)^{-1}(j)) \\ &= \frac{\alpha(i)}{\alpha(j)} \mathbb{E}_{s-1}[v(A_s, A_s^{-1}(j))], \end{aligned}$$

where the second equality follows from the fact that $a^{-1}(i) = (\sigma \circ a)^{-1}(j)$ and the definition of the algorithm ensuring that $\mathbb{P}_{s-1}(A_s = a) = \mathbb{P}_{s-1}(A_s = \sigma \circ a)$. The

last equality follows from the fact that σ is a bijection. Using this and continuing the calculation in Eq. (32.2) shows that

$$\begin{aligned} \text{Eq. (32.2)} &= \frac{\mathbb{E}_{s-1}[v(A_s, A_s^{-1}(i)) - v(A_s, A_s^{-1}(j))]}{\mathbb{E}_{s-1}[v(A_s, A_s^{-1}(i)) + v(A_s, A_s^{-1}(j))]} \\ &= 1 - \frac{2}{1 + \mathbb{E}_{s-1}[v(A_s, A_s^{-1}(i))] / \mathbb{E}_{s-1}[v(A_s, A_s^{-1}(j))]} \\ &\geq 1 - \frac{2}{1 + \alpha(i)/\alpha(j)} \\ &= \frac{\alpha(i) - \alpha(j)}{\alpha(i) + \alpha(j)} = \frac{\Delta_{ij}}{\alpha(i) + \alpha(j)}. \end{aligned}$$

The second part follows from the first since $U_{sji} = -U_{sij}$. \square

The next lemma shows that the failure event occurs with low probability.

LEMMA 32.2 *It holds that $\mathbb{P}(F_n) \leq \delta L^2$.*

Proof The proof follows immediately from Lemma 32.1, the definition of F_n , the union bound over all pairs of actions, and a modification of the Azuma-Hoeffding inequality in Exercise 20.7. \square

LEMMA 32.3 *On the event F_t^c it holds that $(i, j) \notin G_t$ for all $i < j$.*

Proof Let $i < j$ so that $\alpha(i) \geq \alpha(j)$. On the event F_t^c either $N_{sji} = 0$ or

$$S_{sji} - \sum_{u=1}^s \mathbb{E}_{u-1}[U_{uji} \mid U_{uji} \neq 0] |U_{uji}| < \sqrt{2N_{sji} \log\left(\frac{c}{\delta}\sqrt{N_{sji}}\right)} \quad \text{for all } s < t.$$

When i and j are in different blocks in round $u < t$, then $U_{uji} = 0$ by definition. On the other hand, when i and j are in the same block, $\mathbb{E}_{u-1}[U_{uji} \mid U_{uji} \neq 0] \leq 0$ almost surely by Lemma 32.1. Based on these observations,

$$S_{sji} < \sqrt{2N_{sji} \log\left(\frac{c}{\delta}\sqrt{N_{sji}}\right)} \quad \text{for all } s < t,$$

which by the design of TOPRANK implies that $(i, j) \notin G_t$. \square

LEMMA 32.4 *Let $I_{td}^* = \min \mathcal{P}_{td}$ be the most attractive item in \mathcal{P}_{td} . Then on event F_t^c , it holds that $I_{td}^* \leq 1 + \sum_{c < d} |\mathcal{P}_{td}|$ for all $d \in [M_t]$.*

Proof Let $i^* = \min \cup_{c \geq d} \mathcal{P}_{tc}$. Then $i^* \leq 1 + \sum_{c < d} |\mathcal{P}_{td}|$ holds trivially for any $\mathcal{P}_{t1}, \dots, \mathcal{P}_{tM_t}$ and $d \in [M_t]$. Now consider two cases. Suppose that $i^* \in \mathcal{P}_{td}$. Then it must be true that $i^* = I_{td}^*$ and our claim holds. On other hand, suppose that $i^* \in \mathcal{P}_{tc}$ for some $c > d$. Then by Lemma 32.3 and the design of the partition, there must exist a sequence of items i_d, \dots, i_c in blocks $\mathcal{P}_{td}, \dots, \mathcal{P}_{tc}$ such that $i_d < \dots < i_c = i^*$. From the definition of I_{td}^* , $I_{td}^* \leq i_d < i^*$. This concludes our proof. \square

LEMMA 32.5 *On the event F_n^c and for all $i < j$ it holds that*

$$S_{nij} \leq 1 + \frac{6(\alpha(i) + \alpha(j))}{\Delta_{ij}} \log \left(\frac{c\sqrt{n}}{\delta} \right).$$

Proof The result is trivial when $N_{nij} = 0$. Assume from now on that $N_{nij} > 0$. By the definition of the algorithm arms i and j are not in the same block once S_{tij} grows too large relative to N_{tij} , which means that

$$S_{nij} \leq 1 + \sqrt{2N_{nij} \log \left(\frac{c}{\delta} \sqrt{N_{nij}} \right)}.$$

On the event F_n^c and part (a) of Lemma 32.1 it also follows that

$$S_{nij} \geq \frac{\Delta_{ij} N_{nij}}{\alpha(i) + \alpha(j)} - \sqrt{2N_{nij} \log \left(\frac{c}{\delta} \sqrt{N_{nij}} \right)}.$$

Combining the previous two displays shows that

$$\begin{aligned} \frac{\Delta_{ij} N_{nij}}{\alpha(i) + \alpha(j)} - \sqrt{2N_{nij} \log \left(\frac{c}{\delta} \sqrt{N_{nij}} \right)} &\leq S_{nij} \leq 1 + \sqrt{2N_{nij} \log \left(\frac{c}{\delta} \sqrt{N_{nij}} \right)} \\ &\leq (1 + \sqrt{2}) \sqrt{N_{nij} \log \left(\frac{c}{\delta} \sqrt{N_{nij}} \right)}. \end{aligned} \quad (32.3)$$

Using the fact that $N_{nij} \leq n$ and rearranging the terms in the previous display shows that

$$N_{nij} \leq \frac{(1 + 2\sqrt{2})^2 (\alpha(i) + \alpha(j))^2}{\Delta_{ij}^2} \log \left(\frac{c\sqrt{n}}{\delta} \right).$$

The result is completed by substituting this into Eq. (32.3). \square

Proof of Theorem 32.1 The first step in the proof is an upper bound on the expected number of clicks in the optimal list a^* . Fix time t , block \mathcal{P}_{td} , and recall that $I_{td}^* = \min \mathcal{P}_{td}$ is the most attractive item in \mathcal{P}_{td} . Let $k = A_t^{-1}(I_{td}^*)$ be the position of item I_{td}^* and σ be the permutation that exchanges items k and I_{td}^* . By Lemma 32.4, $I_{td}^* \leq k$; and then from Parts (c) and (d) of Assumption 32.1 we have that $v(A_t, k) \geq v(\sigma \circ A_t, k) \geq v(a^*, k)$. Based on this result, the expected number of clicks on I_{td}^* is bounded from below by those on items in a^* ,

$$\begin{aligned} \mathbb{E}_{t-1} [C_{tI_{td}^*}] &= \sum_{k \in \mathcal{I}_{td}} \mathbb{P}_{t-1}(A_t^{-1}(I_{td}^*) = k) \mathbb{E}_{t-1}[v(A_t, k) \mid A_t^{-1}(I_{td}^*) = k] \\ &= \frac{1}{|\mathcal{I}_{td}|} \sum_{k \in \mathcal{I}_{td}} \mathbb{E}_{t-1}[v(A_t, k) \mid A_t^{-1}(I_{td}^*) = k] \geq \frac{1}{|\mathcal{I}_{td}|} \sum_{k \in \mathcal{I}_{td}} v(a^*, k), \end{aligned}$$

where we also used the fact that TOPRANK randomizes within each block to guarantee that $\mathbb{P}_{t-1}(A_t^{-1}(I_{td}^*) = k) = 1/|\mathcal{I}_{td}|$ for any $k \in \mathcal{I}_{td}$. Using this and the design of TOPRANK,

$$\sum_{k=1}^K v(a^*, k) = \sum_{d=1}^{M_t} \sum_{k \in \mathcal{I}_{td}} v(a^*, k) \leq \sum_{d=1}^{M_t} |\mathcal{I}_{td}| \mathbb{E}_{t-1} [C_{tI_{td}^*}] .$$

Therefore, under event F_t^c , the conditional expected regret in round t is bounded by

$$\begin{aligned}
\sum_{k=1}^K v(a^*, k) - \mathbb{E}_{t-1} \left[\sum_{j=1}^L C_{tj} \right] &\leq \mathbb{E}_{t-1} \left[\sum_{d=1}^{M_t} |\mathcal{P}_{td}| C_{tI_{td}^*} - \sum_{j=1}^L C_{tj} \right] \\
&= \mathbb{E}_{t-1} \left[\sum_{d=1}^{M_t} \sum_{j \in \mathcal{P}_{td}} (C_{tI_{td}^*} - C_{tj}) \right] \\
&= \sum_{d=1}^{M_t} \sum_{j \in \mathcal{P}_{td}} \mathbb{E}_{t-1}[U_{tI_{td}^* j}] \\
&\leq \sum_{j=1}^L \sum_{i=1}^{\min\{K, j-1\}} \mathbb{E}_{t-1}[U_{tij}] . \tag{32.4}
\end{aligned}$$

The last inequality follows by noting that $\mathbb{E}_{t-1}[U_{tI_{td}^* j}] \leq \sum_{i=1}^{\min\{K, j-1\}} \mathbb{E}_{t-1}[U_{tij}]$. To see this use part (a) of Lemma 32.1 to show that $\mathbb{E}_{t-1}[U_{tij}] \geq 0$ for $i < j$ and Lemma 32.4 to show that when $I_{td}^* > K$, then neither I_{td}^* nor j are shown to the user in round t so that $U_{tI_{td}^* j} = 0$. Substituting the bound in Eq. (32.4) into the regret leads to

$$R_n \leq nK\mathbb{P}(F_n) + \sum_{j=1}^L \sum_{i=1}^{\min\{K, j-1\}} \mathbb{E}[\mathbb{I}\{F_n^c\} S_{nij}] , \tag{32.5}$$

where we used the fact that the maximum number of clicks over n rounds is nK . The proof of the first part is completed by using Lemma 32.2 to bound the first term and Lemma 32.5 to bound the second. The problem independent bound follows from Eq. (32.5) and by stopping early in the proof of Lemma 32.5 (Exercise 32.6). \square

32.4 Notes

- 1 At no point in the analysis did we use the fact that v is fixed over time. Suppose that v_1, \dots, v_n are a sequence of click-probability functions that all satisfy Assumption 32.1 with the same attractiveness function. The regret in this setting is

$$R_n = \sum_{t=1}^n \sum_{k=1}^K v_t(a^*, k) - \mathbb{E} \left[\sum_{t=1}^n \sum_{i=1}^L C_{ti} \right] .$$

Then the bounds in Theorem 32.1 still hold without changing the algorithm.

- 2 The cascade model is usually formalized in the following more restrictive fashion. Let $\{Z_{ti} : i \in [L], t \in [n]\}$ be a collection of independent Bernoulli random

variables with $\mathbb{P}(Z_{ti} = 1) = \alpha(i)$. Then define K_t as the first item i in the shortlist with $Z_{ti} = 1$:

$$K_t = \min \{k \in [K] : Z_{tA_t(k)} = 1\},$$

where the minimum of an empty set is ∞ . Finally let $C_{ti} = 1$ if and only if $K_t \leq K$ and $A_t(K_t) = i$. This setup satisfies Eq. (32.1), but the independence assumption makes it possible to estimate α without randomization. Notice that in any round t with $K_t \leq K$, all items i with $A_t^{-1}(i) < K_t$ must have been unattractive ($Z_{ti} = 0$) while the clicked item must be attractive ($Z_{ti} = 1$). This fact can be used in combination with standard concentration analysis to estimate the attractiveness. The optimistic policy sorts the L items in decreasing order by their upper confidence bounds and shortlists the first K . When the confidence bounds are derived from Hoeffding's inequality this policy is called CascadeUCB, while the policy that uses Chernoff's lemma is called CascadeKL-UCB. The computational cost of the latter policy is marginally higher than the former, but the improvement is also quite significant because in practice most items have barely positive attractiveness.

- 3 The linear dependence of the regret on L is unpleasant when the number of items is large, which is the case in many practical problems. Like for finite-armed bandits one can introduce a linear structure on the items by assuming that $\alpha(i) = \langle \theta, \phi_i \rangle$ where $\theta \in \mathbb{R}^d$ is an unknown parameter vector and $(\phi_i)_{i=1}^L$ are known feature vectors. This has been investigated in the cascade model by [Zong et al. \[2016\]](#).
- 4 There is an adversarial variant of the cascade model. In the **ranked bandit model** an adversary secretly chooses a sequence of sets S_1, \dots, S_n with $S_t \subseteq [L]$. In each round t the learner chooses $A_t \in \mathcal{A}$ and receives a reward $X_t(A_t)$ where $X_t : \mathcal{A} \rightarrow [0, 1]$ is given by $X_t(a) = \mathbb{I}\{S_t \cap \{a(1), \dots, a(k)\} \neq \emptyset\}$. The feedback is the position of the clicked action, which is $K_t = \min\{k \in [K] : A_t(k) \in S_t\}$. The regret is

$$R_n = \sum_{t=1}^n (X_t(a_*) - X_t(A_t)),$$

where a_* is the optimal ranking in hindsight:

$$a_* = \operatorname{argmin}_{a \in \mathcal{A}} \sum_{t=1}^n X_t(a). \quad (32.6)$$

Notice that this is the same as the cascade model when $S_t = \{i : Z_{ti} = 1\}$.

- 5 A challenge in the ranked bandit model is that solving the offline problem (Eq. 32.6) for known S_1, \dots, S_n is NP-hard. How can one learn when finding an optimal solution to the offline problem is hard? First, hardness only matters if $|\mathcal{A}|$ is large. When L and K are not too large, then exhaustive search is a quite feasible. If this is not an option one may use an approximation algorithm. It turns out that in a certain sense the best one can do is to use a greedy

algorithm, We omit the details, but the highlight is that there exist efficient algorithms such that

$$\mathbb{E} \left[\sum_{t=1}^n X_t(A_t) \right] \geq \left(1 - \frac{1}{e}\right) \max_{a \in \mathcal{A}} \sum_{t=1}^n X_t(a) - O\left(K \sqrt{nL \log(L)}\right).$$

See the article by [Radlinski et al. \[2008\]](#) for more details.

- 6 By modifying the reward function one can also define an adversarial variant of the document-based model. Like before the adversary secretly chooses S_1, \dots, S_n as subsets of $[L]$, but now the reward is

$$X_t(a) = |S_t \cap \{a(1), \dots, a(k)\}|.$$

The feedback is the positions of the clicked items, $S_t \cap \{a(1), \dots, a(k)\}$. For this model there are no computation issues. In fact, problem can be analyzed using a reduction to combinatorial semibandits, which we ask you to investigate in [Exercise 32.3](#).

- 7 The position-based model can also be modelled in the adversarial setting by letting $S_{tk} \subset [L]$ for each $t \in [n]$ and $k \in [K]$. Then defining the reward by

$$X_t(a) = \sum_{k=1}^K \mathbb{I}\{A_t(k) \in S_{tk}\}.$$

Again, the feedback is the positions of the clicked items, $\{k \in [K] : A_t(k) \in S_{tk}\}$. This model can also be tackled using algorithms for combinatorial semibandits ([Exercise 32.4](#)).

32.5 Bibliographic remarks

The policy and analysis presented in this chapter is by the authors and others [[Lattimore et al., 2018](#)]. The most related work is by [Zoghi et al. \[2017\]](#) who assumed a factorization of the click probabilities $v(a, k) = \alpha(a(k))\chi(a, k)$ and then made assumptions on χ . The assumptions made here are slightly less restrictive and the bounds are simultaneously stronger. Some experimental results comparing these algorithms are given by [Lattimore et al. \[2018\]](#). For more information on click models we recommend the survey paper by [Chuklin et al. \[2015\]](#) and article by [Craswell et al. \[2008\]](#). Cascading bandits were first studied by [Kveton et al. \[2015a\]](#), who proposed algorithms based on UCB and KL-UCB and prove finite-time instance-dependence upper bounds and asymptotic lower bounds that match is specific regimes. Around the same time [Combes et al. \[2015\]](#) proposed a different algorithm for the same model that is also asymptotically optimal. The optimal regret has a complicated form and is not given explicitly in all generality. We remarked in the notes that the linear dependence on L is problematic for large L . To overcome this problem [Zong et al. \[2016\]](#) introduce a linear variant where the attractiveness of an item is assumed to be an inner product between an unknown

parameter and a known feature vector. A slightly generalized version of this setup was simultaneously studied by [Li et al. \[2016\]](#), who allowed the features associated with each item to change from round to round. The position-based model is studied by [Lagree et al. \[2016\]](#) who suggest several algorithms and provide logarithmic regret analysis for some of them. Asymptotic lower bounds are also given that match the upper bounds in some regimes. [Katariya et al. \[2016\]](#) study the **dependent click model** introduced by [Guo et al. \[2009\]](#). This differs from the models proposed in this chapter because the reward is not assumed to be the number of clicks and is actually unobserved. We leave the reader to explore this interesting model on their own. The adversarial variant of the ranking problem mentioned in the notes is due to [Radlinski et al. \[2008\]](#). Another related problem is the rank-1 bandit problem where the learner chooses one of L items to place in one of K positions, with all other positions left empty. This model has been investigated by [Katariya et al. \[2017b,a\]](#), who assume the position-based model. The cascade feedback model is also used in a combinatorial setting by [Kveton et al. \[2015c\]](#), but this paper does not have a direct application to ranking.

32.6 Exercises

32.1 Show that the document-based, position-based and cascade models all satisfy Assumption 32.1.

32.2 Most ranking algorithms are based on assigning an attractiveness value to each item and shortlisting the K most attractive items. [Radlinski et al. \[2008\]](#) criticize this approach in their paper as follows:

“The theoretical model that justifies ranking documents in this way is the probabilistic ranking principle [[Robertson, 1977](#)]. It suggests that documents should be ranked by their probability of relevance to the query. However, the optimality of such a ranking relies on the assumption that there are no statistical dependencies between the probabilities of relevance among documents – an assumption that is clearly violated in practice. For example, if one document about jaguar cars is not relevant to a user who issues the query jaguar, other car pages become less likely to be relevant. Furthermore, empirical studies have shown that given a fixed query, the same document can have different relevance to different users [[Teevan et al., 2007](#)]. This undermines the assumption that each document has a single relevance score that can be provided as training data to the learning algorithm. Finally, as users are usually satisfied with finding a small number of, or even just one, relevant document, the usefulness and relevance of a document does depend on other documents ranked higher.”

The optimality criterion [Radlinski et al. \[2008\]](#) had in mind is to present at least one item that the user is attracted to. Do you find this argument convincing? Why or why not?



The probabilistic ranking principle was put forward by [Maron and Kuhns \[1960\]](#). The paper by [Robertson \[1977\]](#) identifies some sufficient conditions under which the principle is valid and also discusses its limitations.

32.3 Frame the adversarial variant of the document-based model in Note 6 as a combinatorial semibandit and use the results in Chapter 30 to prove a bound on the regret of

$$R_n \leq \sqrt{2KLn(1 + \log(L))}.$$

32.4 Adapt your solution to the previous exercise to the position-based model in Note 7 and prove a bound on the regret of

$$R_n \leq K\sqrt{2Ln(1 + \log(L))}.$$

32.5 Prove that if G_t does not contain cycles, then M_t defined in Section 32.2 is well defined and that $\mathcal{P}_{t1}, \dots, \mathcal{P}_{tM_t}$ is a partition of $[L]$.

32.6 Prove the second part of Theorem 32.1.

33 Pure Exploration

All the policies proposed in this book so far have the objective of maximizing the cumulative reward. As a consequence, the policies must carefully balance exploration against exploitation. But what happens if there is no price to be paid for exploring? Imagine, for example, that a researcher has K configurations of a new drug and the budget to test the drugs on n mice. The researcher wants to find the most promising drug configuration for subsequent human trials, but is not concerned with the outcomes for the mice.

Notation

To keep things simple we restrict our attention to K -armed Gaussian bandits with unit variance, but all upper bounds generalize easily to the subgaussian case and with natural modifications to exponential families or other well-behaved distributions. Unless otherwise specified, $\mathcal{E} = \mathcal{E}_{\mathcal{N}}^K(1)$ is the class of Gaussian bandits with unit variance. Since bandits in \mathcal{E} are entirely determined by their mean vectors we identify the two and write $\mu \in \mathcal{E}$ for a Gaussian bandit with mean vector μ . As usual, the learner chooses actions A_1, A_2, \dots with $A_t \in [K]$ and observes rewards X_1, X_2, \dots where

$$X_t = \mu_{A_t} + \eta_t,$$

and η_1, η_2, \dots are independent and identically distributed standard Gaussian random variables. Of course A_t should only depend on $A_1, X_1, \dots, A_{t-1}, X_{t-1}$ and possibly some additional source of randomness. Given $\mu \in \mathbb{R}^K$ let $\Delta_i(\mu) = \max_{j \in [K]} \mu_j - \mu_i$ be the suboptimality gap of the i th arm. For policy π and bandit μ we let $\mathbb{P}_{\mu\pi}$ be the measure on outcomes induced by the interaction of π and μ and $\mathbb{E}_{\mu\pi}[\cdot]$ the expectation with respect to this measure. We also let $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$. Also recall that $T_i(t) = \sum_{s=1}^t \mathbb{I}\{A_s = i\}$ and $\hat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^t \mathbb{I}\{A_s = i\} X_s / T_i(t)$. When the context is obvious we write Δ_i instead of $\Delta_i(\mu)$.

33.1 Simple regret

One way to model the pure exploration problem is to assume a horizon of n rounds. The policy π is expected to output an action A_{n+1} and the loss on bandit

$\mu \in \mathcal{E}$ is the **simple regret**, which is the expected suboptimality gap of the last action:

$$R_n^{\text{SIMPLE}}(\pi, \mu) = \mathbb{E}_{\mu\pi} [\Delta_{A_{n+1}}(\mu)] .$$

In order to get a handle on this new objective we investigate the explore-then-commit algorithm introduced in Chapter 6. Because we only care about choosing a good arm in the final round it makes sense to explore for the first n rounds and then choose the empirically best arm in the last round. Since the commitment is only for the last round, this algorithm is often referred to as the **uniform exploration** policy.

```

1: for  $t = 1, 2, \dots, n$  do
2:   Choose  $A_t = 1 + (t \bmod K)$ 
3: end for
4: Choose  $A_{n+1} = \operatorname{argmax}_{i \in [K]} \hat{\mu}_i(n)$ 
```

Algorithm 18: Explore-then-commit for pure exploration

THEOREM 33.1 *Let π be the policy of Algorithm 18 and $\mu \in \mathcal{E}_{\mathcal{N}}^K(1)$. Then*

$$R_n^{\text{SIMPLE}}(\pi, \mu) \leq \min_{\Delta \geq 0} \left(\Delta + \sum_{i: \Delta_i(\mu) > \Delta} \Delta_i(\mu) \exp \left(-\frac{\lfloor n/K \rfloor \Delta_i(\mu)^2}{4} \right) \right) .$$

Proof Let $\Delta_i = \Delta_i(\mu)$ and $\mathbb{P} = \mathbb{P}_{\mu\pi}$. Assume without loss of generality that $\Delta_1 = 0$ so the first arm is optimal. Let i be a suboptimal arm with $\Delta_i > \Delta$ and observe that $A_{n+1} = i$ implies that $\hat{\mu}_i(n) \geq \hat{\mu}_1(n)$. Now $T_i(n) \geq \lfloor n/K \rfloor$ is not random, so by Theorem 5.1 and Lemma 5.2,

$$\mathbb{P}(\hat{\mu}_i(n) \geq \hat{\mu}_1(n)) = \mathbb{P}(\hat{\mu}_i(n) - \hat{\mu}_1(n) \geq 0) \leq \exp \left(-\frac{\lfloor n/K \rfloor \Delta_i^2}{4} \right) .$$

The definition of the simple regret yields

$$R_n^{\text{SIMPLE}}(\pi, \mu) = \sum_{i=1}^K \Delta_i \mathbb{P}(A_{n+1} = i) \leq \Delta + \sum_{i: \Delta_i > \Delta} \Delta_i \mathbb{P}(A_{n+1} = i) .$$

The proof is completed by taking the minimum over all $\Delta \geq 0$. \square

The theorem highlights some important differences between the simple regret and the cumulative regret. If μ is fixed and n tends to infinity, then the simple regret converges to zero exponentially fast. On the other hand, if n is fixed and μ is allowed to vary, then we are in a worst-case regime. Theorem 33.1 can be used to derive a bound in this case by choosing $\Delta = 2\sqrt{\log(K)/\lfloor n/K \rfloor}$, which after a short algebraic calculation shows there exists a universal constant $C > 0$ such that

$$R_n^{\text{SIMPLE}}(\text{ETC}, \mu) \leq C \sqrt{\frac{K \log(K)}{n}} \quad \text{for all } \mu \in \mathcal{E} . \quad (33.1)$$

In Exercise 33.1 we ask you to use the techniques of Chapter 15 to prove that for all policies there exists a bandit $\mu \in \mathcal{E}$ such that $R_n^{\text{SIMPLE}}(\pi, \mu) \geq C\sqrt{K/n}$ for some universal constant $C > 0$. It turns out the logarithmic dependence on K in Eq. (33.1) is tight for ETC (Exercise 33.2), but there exists another policy for which the simple regret matches the aforementioned lower bound up to constant factors. There are several ways to do this, but the most straightforward is via a reduction from algorithms designed for minimizing cumulative regret.

THEOREM 33.2 *Let π be a policy for which the $(n + 1)$ th action is chosen randomly with $\mathbb{P}(A_{n+1} = i \mid \mathcal{F}_n) = T_i(n)/n$, then its simple regret satisfies*

$$R_n^{\text{SIMPLE}}(\pi, \mu) = \frac{R_n(\pi, \mu)}{n},$$

where $R_n(\pi, \mu)$ is the cumulative regret of policy π when executed on bandit μ .

Proof Using the definition of the regret.

$$R_n(\pi, \mu) = n\mathbb{E} \left[\sum_{i=1}^K \Delta_i \frac{T_i(n)}{n} \right] = n\mathbb{E} [\mathbb{E} [\Delta_{A_{n+1}} \mid \mathcal{F}_n]] = nR_n^{\text{SIMPLE}}(\pi, \mu),$$

where the first equality follows from the definition of the cumulative regret, the third from the definition of the policy in the $(n + 1)$ th round and the last the definition of the simple regret. \square

The theorem raises our hopes that policies designed for minimizing the cumulative regret might also have well-behaved simple regret. Unfortunately this is only true in the intermediate regimes where the best arm is hard to identify. Policies with small cumulative regret spend most of their time playing the optimal arm and play suboptimal arms just barely enough to ensure they are not optimal. In pure exploration this leads to a highly suboptimal policy for which the simple regret is asymptotically polynomial rather than exponential.

33.2 Best arm identification

Let $\delta \in (0, 1)$ be a known confidence level. The objective in **fixed confidence best arm identification** is to design a policy π and \mathcal{F}_t -stopping time τ such that $\mathbb{E}_{\mu\pi}[\tau]$ is as small as possible while ensuring that

$$\mathbb{P}_{\mu\pi}(\Delta_{A_\tau+1}(\mu) > 0) \leq \delta \quad \text{for all } \mu \in \mathcal{E}. \quad (33.2)$$

Like the cumulative regret, minimizing $\mathbb{E}_{\mu\pi}[\tau]$ is a multi-objective criteria and it is not immediately clear that the same policy and stopping rule should minimize $\mathbb{E}_{\mu\pi}[\tau]$ for all $\mu \in \mathcal{E}$ simultaneously. Conveniently, however, the condition that the policy and stopping rule must satisfy Eq. (33.2) plays the role of the consistency assumption in the asymptotic lower bounds in Chapter 16 and for small δ there is a single policy and stopping rule that essentially minimizes $\mathbb{E}_{\mu\pi}[\tau]$ for all μ simultaneously.

Lower bound

We start with the lower bound, which serves as a target for the upper bound to follow. For $\mu \in \mathcal{E}$ define $i^*(\mu) = \operatorname{argmax}_{i \in [K]} \mu_i$ to be the set of optimal arms and

$$\mathcal{E}_{\text{alt}}(\mu) = \{\mu' \in \mathcal{E} : i^*(\mu') \cap i^*(\mu) = \emptyset\},$$

which is the set of Gaussian bandits with different optimal arms than μ .

THEOREM 33.3 *Let $\delta \in (0, 1)$ and suppose that π is a policy and τ a stopping time such that for all $\mu \in \mathcal{E}$ with a unique optimal arm, $\mathbb{P}_{\mu\pi}(A_{\tau+1} \notin i^*(\mu)) \leq \delta$. Then $\mathbb{E}_{\mu\pi}[\tau] \geq c^*(\mu) \log\left(\frac{4}{\delta}\right)$, where*

$$c^*(\mu)^{-1} = \sup_{\alpha \in \Delta^{K-1}} \left(\inf_{\mu' \in \mathcal{E}_{\text{alt}}(\mu)} \left(\sum_{i=1}^K \alpha_i D(\mathcal{N}(\mu_i, 1), \mathcal{N}(\mu'_i, 1)) \right) \right). \quad (33.3)$$

Proof Let $\mu' \in \mathcal{E}_{\text{alt}}(\mu)$. By assumption we have $\mathbb{P}_{\mu\pi}(A_{\tau+1} \notin i^*(\mu)) \leq \delta$ and $\mathbb{P}_{\mu'\pi}(A_{\tau+1} \notin i^*(\mu')) \leq \delta$. The high probability Pinsker's inequality (Theorem 14.2) and the stopping time version of Lemma 15.1 (see Exercise 15.6) show that for any \mathcal{F}_τ -measurable event E ,

$$\mathbb{P}_{\mu\pi}(E) + \mathbb{P}_{\mu'\pi}(E^c) \geq \frac{1}{2} \exp \left(- \sum_{i=1}^K \mathbb{E}_{\mu\pi}[T_i(\tau)] D(\mathcal{N}(\mu_i, 1), \mathcal{N}(\mu'_i, 1)) \right).$$

Choosing $E = \mathbb{I}\{A_{\tau+1} \notin i^*(\mu)\}$ leads to

$$\begin{aligned} 2\delta &\geq \mathbb{P}_{\mu\pi}(A_{\tau+1} \notin i^*(\mu)) + \mathbb{P}_{\mu'\pi}(A_{\tau+1} \notin i^*(\mu')) \\ &\geq \frac{1}{2} \exp \left(- \sum_{i=1}^K \mathbb{E}_{\mu\pi}[T_i(\tau)] D(\mathcal{N}(\mu_i, 1), \mathcal{N}(\mu'_i, 1)) \right). \end{aligned} \quad (33.4)$$

Using the definition of $c^*(\mu)$ and the above display we have

$$\begin{aligned} \frac{\mathbb{E}_{\mu\pi}[\tau]}{c^*(\mu)} &= \mathbb{E}_{\mu\pi}[\tau] \sup_{\alpha \in \Delta^{K-1}} \inf_{\mu' \in \mathcal{E}_{\text{alt}}(\mu)} \sum_{i=1}^K \alpha_i D(\mathcal{N}(\mu_i, 1), \mathcal{N}(\mu'_i, 1)) \\ &\geq \mathbb{E}_{\mu\pi}[\tau] \inf_{\mu' \in \mathcal{E}_{\text{alt}}(\mu)} \sum_{i=1}^K \frac{\mathbb{E}_{\mu\pi}[T_i(\tau)]}{\mathbb{E}_{\mu\pi}[\tau]} D(\mathcal{N}(\mu_i, 1), \mathcal{N}(\mu'_i, 1)) \\ &= \inf_{\mu' \in \mathcal{E}_{\text{alt}}(\mu)} \sum_{i=1}^K \mathbb{E}_{\mu\pi}[T_i(\tau)] D(\mathcal{N}(\mu_i, 1), \mathcal{N}(\mu'_i, 1)) \geq \log\left(\frac{4}{\delta}\right), \end{aligned} \quad (33.5)$$

where the last inequality follows from Eq. (33.4). Rearranging completes the proof. \square

We will shortly show that the lower bound is tight asymptotically as δ tends to zero, but first it is worth examining the value of $c^*(\mu)$. Suppose that $\alpha^*(\mu) \in \Delta^{K-1}$ satisfies

$$c^*(\mu)^{-1} = \inf_{\mu' \in \mathcal{E}_{\text{alt}}(\mu)} \sum_{i=1}^K \alpha_i^*(\mu) D(\mathcal{N}(\mu_i, 1), \mathcal{N}(\mu'_i, 1)).$$

A few observations about this optimization problem:

- (a) Provided that μ has a unique optimal arm, then the value of $\alpha^*(\mu)$ is unique. Uniqueness continues to hold when \mathcal{E} is an unstructured bandit with distributions from an exponential family.
- (b) The inequality in Eq. (33.5) is tightest when $\mathbb{E}_{\mu\pi}[T_i(\tau)]/\mathbb{E}_{\mu\pi}[\tau] = \alpha_i^*(\mu)$, which shows a policy can only match the lower bound by playing arm i exactly in proportion to $\alpha_i^*(\mu)$ in the limit as δ tends to zero.
- (c) When $\mathcal{E} = \mathcal{E}_N^2(1)$ and $\mu \in \mathcal{E}$ has a unique optimal arm, then

$$\begin{aligned} c^*(\mu)^{-1} &= \frac{1}{2} \sup_{\alpha \in [0,1]} \inf_{\mu' \in \mathcal{E}_{\text{alt}}(\mu)} (\alpha(\mu_1 - \mu'_1)^2 + (1-\alpha)(\mu_2 - \mu'_2)^2) \\ &= \frac{1}{2} \sup_{\alpha \in [0,1]} ((1-\alpha)^2 + \alpha^2) (\mu_1 - \mu_2)^2 = \frac{1}{4} (\mu_1 - \mu_2)^2. \end{aligned}$$

In this case we observe that $\alpha_1^*(\mu) = \alpha_2^*(\mu) = 1/2$.

Policy, stopping rule and upper bounds

Both the stopping rule and policy are derived almost directly by the insights derived from the lower bound. For the policy we would like it to choose action i in proportion to $\alpha_i^*(\mu)$, which must be estimated from data. The stopping rule is motivated by recalling from the proof of Theorem 33.3 that for all $\mu' \in \mathcal{E}_{\text{alt}}(\mu)$,

$$\begin{aligned} \mathbb{P}_{\mu\delta}(A_{\tau_\delta+1} \notin i^*(\mu)) + \mathbb{P}_{\mu'\delta}(A_{\tau_\delta+1} \notin i^*(\mu')) &\geq \frac{1}{2} \exp(-D(\mathbb{P}_{\mu\delta}, \mathbb{P}_{\mu'\delta})) \quad (33.6) \\ &= \frac{1}{2} \exp\left(-\sum_{i=1}^K \mathbb{E}[T_i(\tau_\delta)] D(\mathcal{N}(\mu_i, 1), \mathcal{N}(\mu'_i, 1))\right). \end{aligned}$$

If the inequality is tight, then we might guess that a reasonable stopping rule might be the first round t when

$$\sum_{i=1}^K T_i(t) D(\mathcal{N}(\mu_i, 1), \mathcal{N}(\mu'_i, 1)) \gtrsim \log\left(\frac{1}{\delta}\right).$$

There are two problems: (a) μ is unknown, so the expression cannot be evaluated and (b) we have replaced the expected pull counts with their realizations, which may invalidate the expression. Still, let us persevere. To deal with the first problem we can try replacing μ by its estimate $\hat{\mu}(t)$. Then let

$$\begin{aligned} Z_t &= \inf_{\mu' \in \mathcal{E}_{\text{alt}}(\hat{\mu}(t))} \sum_{i=1}^K T_i(t) D(\mathcal{N}(\hat{\mu}_i(t), 1), \mathcal{N}(\mu'_i, 1)) \\ &= \frac{1}{2} \inf_{\mu' \in \mathcal{E}_{\text{alt}}(\hat{\mu}(t))} \sum_{i=1}^K T_i(t) (\hat{\mu}_i(t) - \mu'_i)^2. \end{aligned}$$

We will show there exists a choice of $\beta_t(\delta)$ such that if $\tau_\delta = \min\{t : Z_t \geq \beta_t(\delta)\}$, then the empirically optimal arm at τ_δ is the best arm with probability at least $1 - \delta$. As we remarked earlier, if the policy is to match the lower bound it should

play arm i approximately in proportion to $\alpha_i^*(\mu)$. This suggests estimating $\alpha^*(\mu)$ by $\hat{\alpha}(t) = \alpha^*(\hat{\mu}(t))$ and then playing the arm for which $T_i(t)/\hat{\alpha}_i(t)$ is minimized. If $\hat{\alpha}(t)$ is inaccurate, then perhaps the samples collected will not allow the algorithm to improve its estimates. To overcome this last challenge the policy includes enough forced exploration to ensure that eventually $\hat{\alpha}(t)$ converges to $\alpha^*(\mu)$ with high probability. Combining all these ideas leads to the Track-and-Stop policy (Algorithm 19).

```

1: Input  $\delta$  and  $\beta_t(\delta)$ 
2: Choose each arm once
3: while  $Z_t \leq \beta_t(\delta)$  do
4:   if  $\operatorname{argmin}_{i \in [K]} T_i(t-1) \leq \sqrt{t}$  then
5:     Choose  $A_t = \operatorname{argmin}_{i \in [K]} T_i(t-1)$ 
6:   else
7:     Choose  $A_t = \operatorname{argmin}_{i \in [K]} \frac{T_i(t-1)}{\hat{\alpha}_i^*(t-1)}$ 
8:   end if
9: end while
```

Algorithm 19: Track-and-Stop

THEOREM 33.4 Let π_δ and τ_δ be the policy/stopping rule of Algorithm 19. There exists a choice of $\beta_t(\delta)$ such that for all $\mu \in \mathcal{E}$ with $|i^*(\mu)| = 1$ it holds that

$$\lim_{\delta \rightarrow 0} \frac{\mathbb{E}_{\mu \pi_\delta}[\tau_\delta]}{\log(1/\delta)} = c^*(\mu).$$

Furthermore, $\mathbb{P}_{\mu \pi_\delta}(i^*(\hat{\mu}_{\tau_\delta}) \neq i^*(\mu)) \leq \delta$.

The proof takes a little work. First we show the stopping rule is sound in the sense that indeed the algorithm outputs the optimal arm with probability at least $1 - \delta$.

LEMMA 33.1 Let $f : [K, \infty) \rightarrow \mathbb{R}$ be given by $f(x) = \exp(K-x)(x/K)^K$ and $\beta_t(\delta) = K \log(t^2 + t) + f^{-1}(\delta)$. Then for $\tau = \min\{t : Z_t \geq \beta_t(\delta)\}$ it holds that $\mathbb{P}(i^*(\hat{\mu}(\tau)) \neq i^*(\mu)) \leq \delta$.



Basic calculus shows that f is monotone decreasing on $[K, \infty)$ so the inverse is well defined. In fact the inverse has a closed form solution in terms of the Lambert W function. By staring at the form of f one can check that $\lim_{\delta \rightarrow 0} f^{-1}(\delta)/\log(1/\delta) = 1$ or equivalently that $f^{-1}(\delta) = (1 + o(1)) \log(1/\delta)$.

Proof of Lemma 33.1 Assume that $\mu_1 = \max_i \mu_i$. By the definition of τ , if $\mu \in \mathcal{E}_{\text{alt}}(\hat{\mu}(\tau))$, then

$$\frac{1}{2} \sum_{i=1}^K T_i(\tau) (\hat{\mu}_i(\tau) - \mu_i)^2 \geq \beta_\tau(\delta).$$

Using the definition of $\mathcal{E}_{\text{alt}}(\hat{\mu}(\tau))$ yields

$$\mathbb{P}(1 \neq i^*(\hat{\mu}(\tau))) = \mathbb{P}(\mu \in \mathcal{E}_{\text{alt}}(\hat{\mu}(\tau))) \leq \mathbb{P}\left(\frac{1}{2} \sum_{i=1}^K T_i(\tau)(\hat{\mu}_i(\tau) - \mu_i)^2 \geq \beta_\tau(\delta)\right).$$

Then apply Lemma 33.2 and Proposition 33.1 from Section 33.2.1. \square

Below we sketch the proof of Theorem 33.4. A more complete outline is given in Exercise 33.6.

Proof sketch of Theorem 33.4 Lemma 33.1 shows that the stopping procedure and selection rule of Track-and-Stop are valid in the sense that the probability of the arm selected being suboptimal is at most δ . It remains to control the expectation of the stopping time. The intuition is straightforward. As more samples are collected we expect that $\hat{\alpha}(t) \approx \alpha^*(\mu)$ and $\hat{\mu}(t) \approx \mu$ and

$$Z_t = \inf_{\tilde{\mu} \in \mathcal{E}_{\text{alt}}(\hat{\mu}(t))} \sum_{i=1}^K \frac{T_i(t)(\hat{\mu}_i(t) - \tilde{\mu}_i)^2}{2} \approx \inf_{\tilde{\mu} \in \mathcal{E}_{\text{alt}}(\mu)} \sum_{i=1}^K \frac{\alpha_i^*(\mu)(\mu_i - \tilde{\mu}_i)^2}{2} = \frac{t}{c^*(\mu)}.$$

Provided the approximation is reasonably accurate, the algorithm should halt once

$$\frac{t}{c^*(\mu)} \geq \beta_t(\delta) = (1 + o(1)) \log(1/\delta),$$

which occurs once $t \geq (1 + o(1))c^*(\mu) \log(1/\delta)$. \square

33.2.1 Concentration

The first concentration theorem follows from Corollary 5.1 and a union bound.

LEMMA 33.2 *Let X_1, X_2, \dots be a sequence of independent Gaussian random variables with mean μ and unit variance. Let $\hat{\mu}_n = \frac{1}{n} \sum_{t=1}^n X_t$. Then*

$$\mathbb{P}\left(\exists n \in \mathbb{N}^+ : \frac{n}{2}(\hat{\mu}_n - \mu)^2 \geq \log(1/\delta) + \log(n(n+1))\right) \leq \delta.$$

PROPOSITION 33.1 *Let $g : \mathbb{N} \rightarrow \mathbb{R}$ be monotone nondecreasing and for each $i \in [K]$ let S_{i1}, S_{i2}, \dots be an infinite sequence of random variables such that for all $\delta \in (0, 1)$,*

$$\mathbb{P}(\exists s \in \mathbb{N} : S_{is} \geq g(s) + \log(1/\delta)) \leq \delta.$$

Then provided that $(S_i)_{i=1}^K$ are independent and $x \geq 0$,

$$\mathbb{P}\left(\exists s \in \mathbb{N}^K : \sum_{i=1}^K S_{is_i} \geq K g\left(\sum_{i=1}^K s_i\right) + x\right) \leq \left(\frac{x}{K}\right)^K \exp(K - x).$$

Proof For $i \in [d]$ let $W_i = \max\{w \in [0, 1] : S_{is} < g(s) + \log(1/w) \text{ for all } s \in \mathbb{N}\}$. Then for any $s \in \mathbb{N}^d$,

$$\sum_{i=1}^d S_{is_i} \leq \sum_{i=1}^d g(s_i) + \sum_{i=1}^d \log(1/W_i) \leq dg\left(\sum_{i=1}^d s_i\right) + \sum_{i=1}^d \log(1/W_i).$$

By assumption $(W_i)_{i=1}^d$ are independent and satisfy $\mathbb{P}(W_i \leq x) \leq x$ for all $x \in [0, 1]$. The proof is completed by using Exercise 5.18. \square

33.3 Best arm identification with a budget

The setting in the previous section is called the fixed confidence version of best arm identification because the learner should minimize the exploration time in order to satisfy a constraint on the confidence. In the fixed budget variant the learner is given a constraint on the horizon and should minimize the probability of choosing a suboptimal arm.

This reframing of the problem makes algorithm design and analysis a little more nuanced and the results are not as clean. A naive option would be to use the explore-then-commit policy, but as discussed in Section 33.1 this approach leads to poor results when the suboptimality gaps are not close. To overcome this problem the Sequential Halving algorithm divides the budget into $L = \lceil \log_2(K) \rceil$ phases. In the first phase the algorithm chooses each arm $\lfloor n/(KL) \rfloor$ times. The bottom half of the arms are eliminated and the process is repeated.

```

1: Input  $n$  and  $K$ 
2: Set  $L = \lceil \log_2(K) \rceil$  and  $\mathcal{A}_1 = [K]$ .
3: for  $\ell = 1, \dots, L$  do
4:   Let  $T_\ell = \left\lfloor \frac{n}{L|\mathcal{A}_\ell|} \right\rfloor$ .
5:   Choose each arm in  $\mathcal{A}_\ell$  exactly  $T_\ell$  times
6:   For each  $i \in \mathcal{A}_\ell$  compute  $\hat{\mu}_i^\ell$  as the empirical mean of arm  $i$  based on the
      last  $T_\ell$  samples
7:   Let  $\mathcal{A}_{\ell+1}$  contain the top  $\lceil |\mathcal{A}_\ell|/2 \rceil$  arms in  $\mathcal{A}_\ell$ 
8: end for
9: Output the arm in  $\mathcal{A}_{L+1}$ 

```

Algorithm 20: Sequential Halving

Let $\mu \in \mathcal{E}$ and assume that $\mu_1 > \mu_2 \geq \dots \geq \mu_K$. Define $H_1(\mu)$ and $H_2(\mu)$ by

$$H_1(\mu) = \sum_{i=2}^K \frac{1}{\Delta_i^2} \quad H_2(\mu) = \max_{i>1} \frac{i}{\Delta_i^2}.$$

For bandits where the arms are not in order the value of $H_i(\mu)$ is defined as above after permuting the arms. The quantity $H_2(\mu)$ looks a bit unusual, but we will see it arises quite naturally in the analysis. The following also holds:

$$H_2(\mu) \leq H_1(\mu) \leq \frac{H_2(\mu)}{1 + \log(K)}. \quad (33.7)$$

THEOREM 33.5 *If $\mu \in \mathcal{E}$ and π is sequential halving, then*

$$\mathbb{P}_{\mu\pi}(\Delta_{A_{n+1}} > 0) \leq 3 \log_2(K) \exp\left(-\frac{n}{16H_2(\mu) \log_2(K)}\right).$$

In Exercise 33.7 the reader is guided through the proof of this theorem. Let's see how the bound compares to explore-then-commit, which is the same as Algorithm 18. Like in the proof of Theorem 33.1, the probability that ETC selects a suboptimal arm is easily controlled using Theorem 5.1 and Lemma 5.2:

$$\mathbb{P}_{\mu,ETC}(\Delta_{A_{n+1}} > 0) \leq \sum_{i=2}^K \mathbb{P}(\hat{\mu}_i(n) \geq \hat{\mu}_1(n)) \leq \sum_{i=2}^K \exp\left(-\frac{\lfloor n/K \rfloor \Delta_i^2}{4}\right).$$

Suppose that $\Delta = \Delta_2 = \Delta_K$ so that all suboptimal arms have the same suboptimality gap. Then $H_2 = K/\Delta^2$ and terms in the exponent for sequential halving and ETC are $O(n\Delta^2/(K \log K))$ and $O(n\Delta^2/K)$ respectively, which means that ETC is actually moderately better than sequential halving, at least if n is sufficiently large. On the other hand, if Δ_2 is small, but $\Delta_i = 1$ for all $i > 2$, then $H_2 = O(1/\Delta_2^2)$ and the exponents are $O(n\Delta^2)$ and $O(n\Delta^2/K)$ respectively and sequential halving is significantly better. The reason for the disparity is the non-adaptivity of ETC, which wastes many samples on arms $i > 2$. On the other hand, with high probability the sequential halving algorithm spends one quarter of its budget sampling from arm two.

33.4 Notes

- 1 We mentioned briefly that algorithms with logarithmic cumulative regret are not well suited for pure exploration. Suppose that π is a policy such that for each $i \in [K]$ it holds that

$$\mathbb{E}_{\mu\pi}[T_i(n)] = \frac{2}{\Delta_i^2} \log(n) + o(\log(n)) \quad \text{for all } \mu \in \mathcal{E}.$$

We showed that such policies exist in Chapter 8 and that one cannot do better in Chapter 16. Let $\mu \in \mathcal{E}$ be a bandit for which there is a unique optimal arm and let $\mu' \in \mathcal{E}_{\text{alt}}(\mu)$ be the alternative bandit that has the same mean rewards as μ for all arms except $\mu'_i = \mu_i + (1 + \varepsilon)\Delta_i$. Then by Theorem 14.2 and Lemma 15.1,

$$\begin{aligned} \mathbb{P}_{\mu\pi}(A_{n+1} \neq 1) + \mathbb{P}_{\mu'\pi}(A_{n+1} \neq i) &\geq \frac{1}{2} \exp(-D(\mathbb{P}_{\mu\pi}, \mathbb{P}_{\mu'\pi})) \\ &\geq \frac{1}{2} \exp(-(log(n) + o(log(n)))(1 + \varepsilon)^2) = \frac{1}{2} \left(\frac{1}{n}\right)^{(1+o(1))(1+\varepsilon)^2}. \end{aligned}$$

This shows that using an asymptotically optimal policy for cumulative regret minimization leads to a best arm identification policy for which the probability of selecting a suboptimal arm decays only polynomially with n . Note that here

we did not make any restrictions on the selection rule that determines A_{n+1} , only that the first n samples were collected by an asymptotically optimal regret minimizer.

- 2 Although there is no exploration/exploitation dilemma in the pure exploration setting, there is still an ‘exploration dilemma’ in the sense that the optimal exploration policy depends on an unknown quantity. This means the policy must balance (to some extent) the number of samples dedicated to learning the how to explore relative to those actually exploring.
- 3 The forced exploration in the Track-and-Stop algorithm is good enough for asymptotic optimality, but the fact that the proof would go through with almost any sublinear amount of exploration should cause a little unease. We do not currently know of a principled way to tune the amount of forced exploration, or indeed if there is better algorithm design for best arm identification.
- 4 The choice of $\beta_t(\delta)$ significantly influences the practical performance of Track-and-Stop. We believe the analysis given here is mostly tight except that the naive concentration bound given in Lemma 33.2 can be improved significantly.
- 5 Perhaps the most practical setup in pure exploration has not yet received any attention, which is upper and lower instance-dependent bounds on the simple regret. Even better, an analysis of the distribution of $\Delta_{A_{n+1}}$.

33.5 Bibliographical remarks

The study of pure exploration for bandits seems to have been first studied by Even-Dar et al. [2002], Mannor and Tsitsiklis [2004], Even-Dar et al. [2006] in the ‘Probability Approximately Correct’ setting where the objective is to find an ε -optimal with as few samples as possible. After a dry spell the field was restarted by Bubeck et al. [2009], Audibert and Bubeck [2010b]. Asymptotically optimal algorithms in the fixed confidence setting of Section 33.2 were introduced at the same conference by Garivier and Kaufmann [2016] and Russo [2016], both of which are heartily recommended. The algorithm and analysis presented here is based on the first of these two articles, which also provides results for exponential families as well as in-depth intuition and historical background. The stopping rule used by Garivier and Kaufmann [2016] is inspired by similar rules by Chernoff [1959]. The sequential halving algorithm is by Karnin et al. [2013]. Besides this there have been many other approaches: Jamieson and Nowak [2014]. The negative result showing that policies for minimizing the cumulative regret do not explore enough in the pure exploration setting is due to Bubeck et al. [2009]. For lower bounds in the fixed budget problem we refer the reader to the recent paper by Carpentier and Locatelli [2016]. Pure exploration has recently become a hot topic and is expanding beyond the finite-armed case. For example, to linear bandits [Soare et al., 2014] and continuous armed bandits and tree search [Garivier et al., 2016a, Huang et al., 2017a].

Continuous-armed case: Munos [2011], Valko et al. [2013a] and more.

33.6 Exercises

33.1 Show there exists a universal constant $C > 0$ such that for all $n \geq K > 1$ and all policies π there exists a $\mu \in \mathcal{E}$ such that

$$R_n^{\text{SIMPLE}}(\pi, \mu) \geq C \sqrt{\frac{K}{n}}.$$

33.2 Show there exists a universal constant $C > 0$ such that for all $n \geq K > 1$ there exists a $\mu \in \mathcal{E}$ such that

$$R_n^{\text{SIMPLE}}(\text{ETC}, \mu) \geq C \sqrt{\frac{K \log(K)}{n}}.$$

33.3 Prove both inequalities in Eq. (33.7).

33.4 This exercise is about designing (ε, δ) -PAC algorithms.

- (a) For each $\varepsilon > 0$ and $\delta \in (0, 1)$ and number of arms $K > 1$ design a policy π and stopping time τ such that for all $\mu \in \mathcal{E}$,

$$\mathbb{P}_{\mu\pi}(\Delta_{A_\tau} \geq \varepsilon) \leq \delta \quad \text{and} \quad \mathbb{E}_{\mu\pi}[\tau] \leq \frac{CK}{\varepsilon^2} \log\left(\frac{K}{\delta}\right),$$

for universal constant $C > 0$.

- (b) It turns out the logarithmic dependence on K can be eliminated. Design a policy π and stopping time τ such that for all $\mu \in \mathcal{E}$,

$$\mathbb{P}_{\mu\pi}(\Delta_{A_\tau} \geq \varepsilon) \leq \delta \quad \text{and} \quad \mathbb{E}_{\mu\pi}[\tau] \leq \frac{CK}{\varepsilon^2} \log\left(\frac{1}{\delta}\right).$$

- (c) Prove a lower bound showing that the bound in part (b) is tight up to constant factors in the worst case.



Part (b) of the above exercise is a challenging problem. The simplest approach is to use an elimination algorithm that operates in phases where at the end of each phase the bottom half of the arms (in terms of their empirical estimates) are eliminated. For details see the paper by Even-Dar et al. [2002].

33.5 Let $K = 2$ and suppose a bandit policy π has a cumulative regret of $R_{n-1}(\pi, \mu) \leq C_n(\mu) \log(n)$ where $C_n : \mathcal{E} \rightarrow [0, \infty)$ is an instance-dependent constant. Suppose this policy is run for $n-1$ steps and subsequently the empirically best arm is played.

- (a) Show there exists a $\mu' \in \mathcal{E}$ such that

$$\mathbb{P}_{\mu\pi}(\Delta_{A_n} > 0) + \mathbb{P}_{\mu'\pi}(\Delta_{A_n} > 0) \geq \frac{1}{2} \exp\left(-\frac{1}{2} C_n(\mu) \Delta(\mu) \log(n)\right),$$

where $\Delta(\mu) = |\mu_1 - \mu_2|$.

- (b) Suppose that π is asymptotically optimal in the sense that $\lim_{n \rightarrow \infty} C_n(\mu) = 2/\Delta(\mu)$. Show that

$$\lim_{n \rightarrow \infty} \sup_{\mu \in \mathcal{E}} n \mathbb{P}_{\mu\pi}(\Delta_{A_n} > 0) \geq 1.$$

33.6 In this exercise you will complete the proof of Theorem 33.4. Define

$$\begin{aligned}\Phi(\mu, \alpha) &= \inf_{\tilde{\mu} \in \mathcal{E}_{\text{alt}}(\mu)} \sum_{i=1}^K \alpha_i (\mu_i - \tilde{\mu}_i)^2. \\ M(\varepsilon) &= \min\{t : \sup_{s \geq t} |\Phi(\hat{\mu}_t, T(t)/t) - \Phi(\mu, \alpha^*)| \leq \varepsilon\}. \\ t^*(\varepsilon, \delta) &= \min\{t : \sup_{s \geq t} s(\Phi(\mu, \alpha^*) - \varepsilon) - \beta_s(\delta) \geq 0\}.\end{aligned}$$

Let $F_{t,\varepsilon}$ be the event that $\|\hat{\mu}_t - \mu\|_\infty \leq \varepsilon$.

- (a) Assume that μ has a unique optimal arm. Show that Φ is continuous at μ .
(b) Show that if $\cup_{s \geq t} F_{s,\varepsilon}$, then

$$\|\alpha_s - \alpha_s^*\|_\infty \leq 3\varepsilon.$$

- (c) Show that

$$\|T(t)/t - \alpha_t^*\|_\infty$$

- (d) Let $\varepsilon > 0$ and $t_\delta^* = \min\{t : tc^*(\mu) \geq \beta_t(\delta)\}$. Show that

$$\mathbb{E}[\tau_\delta] \leq (1 + \varepsilon)t_\delta^* + \sum_{t=\lceil(1+\varepsilon)t_\delta^*\rceil}^{\infty} \mathbb{P}(|\Phi(\hat{\mu}_t, \hat{\alpha}_t) - \Phi(\mu, \alpha^*)| \geq \varepsilon c^*(\mu)).$$

- (e) Use the continuity of Φ to show there exists a function $\zeta : [0, \infty) \rightarrow \mathbb{R}$ with $\lim_{\varepsilon \rightarrow 0} \zeta(\varepsilon) = 0$ such that

$$\mathbb{P}(|\Phi(\hat{\mu}_t, \hat{\alpha}_t) - \Phi(\mu, \alpha^*)| \geq c^*(\mu)\varepsilon) \leq \mathbb{P}(\|\hat{\mu}_t - \mu\|_\infty \geq \zeta(\varepsilon)).$$

- (f) Prove that $\mathbb{P}(\|\hat{\mu}_t - \mu\|_\infty \geq \zeta(\varepsilon)) \leq K \exp\left(-\frac{\lfloor \sqrt{t/K} \rfloor \zeta(\varepsilon)^2}{2}\right)$.

- (g) Show that $\lim_{\delta \rightarrow 0} \frac{t_\delta^*}{\log(1/\delta)} = c^*(\mu)$.

- (h) Combine the previous parts to complete the proof of Theorem 33.4 by showing that

$$\limsup_{\delta \rightarrow 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq c^*(\mu).$$



Part (b) is by far the hardest step. Use the forced exploration to prove reasonably fast convergence of $\hat{\mu}_t$ to μ and then continuity arguments. For more details see the article by Garivier and Kaufmann [2016].

33.7 The purpose of this exercise is to prove Theorem 33.5. Assume without loss of generality that $\mu_1 \geq \mu_2 \geq \dots \geq \mu_K$. Given a set $A \subset [K]$ let

$$\text{TOPK}(A, k) = \left\{ i \in [K] : \sum_{j \leq i} \mathbb{I}\{j \in A\} \leq k \right\}$$

be the top k arms in A . To make life easier you may also assume that K is a power of two so that $|\mathcal{A}_\ell| = K2^{1-\ell}$ and $T_\ell = n2^{\ell-1}/\log_2(K)$.

- (a) Prove that $|\mathcal{A}_{L+1}| = 1$.
- (b) Let i be a suboptimal arm in \mathcal{A}_ℓ and suppose that $1 \in \mathcal{A}_\ell$. Show that

$$\mathbb{P}\left(\hat{\mu}_1^\ell \leq \hat{\mu}_i^\ell \mid i \in \mathcal{A}_\ell, 1 \in \mathcal{A}_\ell\right) \leq \exp\left(-\frac{T_\ell \Delta_i^2}{4}\right).$$

- (c) Let $\mathcal{A}'_\ell = \mathcal{A}_\ell \setminus \text{TOPK}(\mathcal{A}_\ell, \lceil |\mathcal{A}_\ell|/4 \rceil)$ be the bottom three quarters of the arms in round ℓ . Show that if the optimal arm is eliminated after the ℓ th phase, then

$$N_\ell = \sum_{i \in \mathcal{A}'_\ell} \mathbb{I}\{\hat{\mu}_i^\ell \geq \hat{\mu}_1^\ell\} \geq \frac{1}{3} |\mathcal{A}'_\ell|.$$

- (d) Let $i_\ell = \min \mathcal{A}'_\ell$ and show that

$$\mathbb{E}[N_\ell \mid \mathcal{A}_\ell] \leq |\mathcal{A}'_\ell| \max_{i \in \mathcal{A}'_\ell} \exp\left(-\frac{\Delta_i^2 n 2^{\ell-1}}{4 \log_2(K)}\right) \leq |\mathcal{A}'_\ell| \exp\left(-\frac{n \Delta_{i_\ell}^2}{16 i_\ell \log_2(K)}\right).$$

- (e) Combine the previous two parts with Markov's inequality to show that

$$\mathbb{P}(1 \notin \mathcal{A}_{\ell+1} \mid 1 \in \mathcal{A}_\ell) \leq 3 \exp\left(-\frac{T \Delta_{i_\ell}^2}{16 \log_2(K) i_\ell}\right).$$

- (f) Join the dots to prove Theorem 33.5.

34 Bayesian Methods

Bayesian methods have been applied to bandits from the very beginning [Thompson, 1933]. At a technical level the difference between Bayesian and frequentist methods is that the Bayesian includes the unknown hypothesis about the world in the probability space. This means a Bayesian can express the likelihood of a hypothesis conditioned on having observed some data. By contrast, a frequentist only has access to the likelihood of the data under fixed hypotheses.

The ability to reason probabilistically about the truth or otherwise of a hypothesis seems like a good thing. The downside is that by including hypotheses in the probability space one is compelled to assign a prior belief about the likelihood of each hypothesis without having observed any data. This is a double edged sword with the potential for poor performance if the prior is not reflected by reality. On the other hand, if the prior beliefs are well-founded it would be strange not to try and use them. The debate between frequentist and Bayesian schools of thought does not interest us greatly. Bayesian approaches to bandits have their strengths and weaknesses and we hope to do them a modicum of justice here.

Let \mathcal{E} be a set of finite-armed stochastic bandits. Recall the regret of policy π in environment $\nu \in \mathcal{E}$ over n rounds is

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

Now let \mathcal{G} be a σ -algebra over \mathcal{E} so that $(\mathcal{E}, \mathcal{G})$ is a measurable space and let \mathbb{Q} be a probability measure on this space called the **prior**. The **Bayesian regret** of policy π is the expected regret of π over all environments in \mathcal{E} with respect to the prior \mathbb{Q} .

$$\text{BR}_n(\pi, \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}} [R_n(\pi, \nu)].$$

Implicit in this definition is the assumption that \mathcal{G} is sufficiently rich that $R_n(\pi, \nu)$ is \mathcal{G} -measurable, which is true for all reasonable choices of \mathcal{G} and policies π . Given a prior and policy the Bayesian regret is just a number. The Bayesian optimal value is $\text{BR}_n^*(\mathbb{Q}) = \inf_{\pi} \text{BR}_n(\pi, \mathbb{Q})$ and the optimal policy is

$$\pi^* = \operatorname{argmin}_{\pi} \text{BR}_n(\pi, \mathbb{Q}). \quad (34.1)$$

In all generality there is no guarantee that the optimal policy exists, but the

positivity of the Bayesian regret ensures that for any $\varepsilon > 0$ there exists a policy π with $\text{BR}_n(\pi, \mathbb{Q}) \leq \text{BR}_n^*(\mathbb{Q}) + \varepsilon$.



The fact that $R_n(\pi, \nu) \geq 0$ for all ν and π means the Bayesian regret is always nonnegative. Perhaps less obviously, the Bayesian regret of the Bayesian optimal policy can be strictly greater than zero (Exercise 34.1).

The Bayesian approach is attractive in several ways. First, once the prior has been chosen, finding the Bayesian optimal policy reduces to an optimization problem. Second, all tradeoff decisions are incorporated into the prior, which prescribes how much we should care about the regret in one bandit relative to another. There are also some disadvantages. From a practical perspective the biggest issue is that solving Eq. (34.1) exactly is usually not computationally tractable. More philosophically there is often no clear way to choose the prior \mathbb{Q} and the choices are often restricted by computational challenges. The computational intractability of finding the exact Bayesian optimal policy makes it a little ironic that approximately Bayesian methods can often be more efficient than frequentist policies. We examine some of these approximate methods in the next chapter. Here we focus on the behaviour of the Bayesian regret and the important settings where Bayesian optimal policies can be computed with reasonable efficiency.



Analyzing the Bayesian regret of an algorithm is strictly less informative than the frequentist regret. By this we mean that a bound on $\text{BR}_n(\pi, \mathbb{Q})$ cannot usually be used to obtain a meaningful bound on $R_n(\pi, \nu)$, while if $R_n(\pi, \nu) \leq f(\nu)$ for a measurable function f , then clearly $\text{BR}_n(\pi, \mathbb{Q}) \leq \mathbb{E}[f(\nu)]$. This is not an argument against using a Bayesian algorithm, but rather an argument to analyze the frequentist regret of Bayesian algorithms.

34.1 Bayesian optimal regret for finite-armed bandits

In the standard finite-armed bandit model the Bayesian optimal policy cannot be computed efficiently. Nevertheless, one can investigate the value of the Bayesian optimal regret by proving upper and lower bounds. One way to upper bound the Bayesian optimal regret is to integrate the frequentist regret bound of one of the algorithms in Part II.

For simplicity we restrict our attention to Bernoulli bandits, but the arguments generalize more broadly. Let $\mathcal{E} = \mathcal{E}_{\mathcal{B}}^K$ be the set of K -armed Bernoulli bandits. Bandits in this class are characterized by their mean vectors so we identify \mathcal{E} with $[0, 1]^K$ and define the prior \mathbb{Q} on $([0, 1]^K, \mathfrak{B}([0, 1]^K))$. The Bayesian optimal regret is necessarily smaller than the minimax regret, which by Theorem 9.1

means that

$$\text{BR}_n^*(\mathbb{Q}) \leq C\sqrt{Kn},$$

where $C > 0$ is a universal constant. The proof of the lower bound in Exercise 15.1 shows that for each n there exists a prior for which

$$\text{BR}_n^*(\mathbb{Q}) \geq c\sqrt{Kn}.$$

where $c > 0$ is a universal constant. In fact one can find a single prior such that this nearly holds for all n . We ask you to prove the following theorem in Exercise 34.3.

THEOREM 34.1 *For any prior \mathbb{Q} ,*

$$\limsup_{n \rightarrow \infty} \frac{\text{BR}_n^*(\mathbb{Q})}{n^{1/2}} = 0.$$

Furthermore, there exists a prior \mathbb{Q} such that for all $\varepsilon > 0$,

$$\liminf_{n \rightarrow \infty} \frac{\text{BR}_n^*(\mathbb{Q})}{n^{1/2-\varepsilon}} = \infty.$$

34.2 Bayesian learning (\dagger)

So far we introduced the Bayesian regret as an average of the frequentist regret and used the results from previous chapters to bound the Bayesian regret. For the rest of the chapter we immerse ourselves in the Bayesian viewpoint by analyzing two special cases where the Bayesian optimal policy can be computed with reasonable accuracy. Before getting into the details we discuss briefly some of the measure-theoretic aspects of Bayesian learning.

Starting gently, suppose you are given a bag containing two marbles. A trustworthy source tells you the bag contains either (a) two white marbles (ww) or (b) a white marble and a black marble (wb). You are allowed to choose a marble from the bag (without looking) and observe its color, which we abbreviate by ‘select white’ (sw) or ‘select black’ (sb). The question is how to update your ‘beliefs’ about the contents of the bag having observed one of the marbles. The Bayesian way to tackle this problem starts by choosing a probability distribution on the space of hypotheses called the prior. This distribution is usually supposed to reflect your prior belief about which hypothesis is more probable. In this case it seems reasonable to choose $\mathbb{P}(\text{ww}) = 1/2$ and $\mathbb{P}(\text{wb}) = 1/2$. The next step is to think about the likelihood of the possible outcomes under each hypothesis. Assuming the bags are well shuffled and you cannot feel the color of a marble these are

$$\mathbb{P}(\text{sw} \mid \text{ww}) = 1 \quad \text{and} \quad \mathbb{P}(\text{sw} \mid \text{wb}) = 1/2.$$

The conditioning here indicates that we are including the hypotheses as part of

the probability space, which is a distinguishing feature of the Bayesian approach. With this formulation we can apply Bayes' law (Eq. 2.1) to show that

$$\begin{aligned}\mathbb{P}(\text{ww} \mid \text{sw}) &= \frac{\mathbb{P}(\text{sw} \mid \text{ww})\mathbb{P}(\text{ww})}{\mathbb{P}(\text{sw})} = \frac{\mathbb{P}(\text{sw} \mid \text{ww})\mathbb{P}(\text{ww})}{\mathbb{P}(\text{sw} \mid \text{ww})\mathbb{P}(\text{ww}) + \mathbb{P}(\text{sw} \mid \text{wb})\mathbb{P}(\text{wb})} \\ &= \frac{1 \times \frac{1}{2}}{1 \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2}} = \frac{2}{3}.\end{aligned}$$

Of course $\mathbb{P}(\text{wb} \mid \text{sw}) = 1 - \mathbb{P}(\text{ww} \mid \text{sw}) = 1/3$. The interpretation of this result is that before observing a white marble your belief in each of the possible scenarios was uniform, but having observed a white marble you now believe the other marble is white with probability 2/3. An alternative calculation shows that $\mathbb{P}(\text{ww} \mid \text{sb}) = 0$, which makes sense because choosing a black marble rules out the hypothesis that the bag contains two white marbles. The conditional distribution $\mathbb{P}(\cdot \mid \text{sw})$ over the hypotheses is called the **posterior** distribution and represents the Bayesian's belief in each hypothesis after selecting a white marble.

Measure-theoretic viewpoint

A more sophisticated approach is necessary when the hypothesis and/or outcome are not discrete. In less mathematical texts the underlying details are often (quite reasonably) swept under the rug for the sake of clarity. Besides the desire for generality there are two reasons not to do this. First, having spent the effort developing the measure-theoretic tools in Chapter 2 it would seem a waste not to use them now. And second, the subtle issues that arise highlight some of the philosophical differences between the Bayesian and frequentist viewpoints that seem worth illuminating.

Let (Θ, \mathcal{G}) be a measurable space called the **parameter space** and (Ω, \mathcal{F}) be a measurable space called the **outcome space**. A prior is a measure \mathbb{Q} on (Θ, \mathcal{G}) and a **hypothesis space** is a probability kernel $\{\mathbb{P}_\theta : \theta \in \Theta\}$ from (Θ, \mathcal{G}) to (Ω, \mathcal{F}) . By the assumption that $\{\mathbb{P}_\theta\}$ is a probability kernel we can define the joint probability measure \mathbb{P} on $(\Theta \times \Omega, \mathcal{G} \otimes \mathcal{F})$ by

$$\mathbb{P}(\theta \in A, \omega \in B) = \int_A \mathbb{P}_\theta(B) d\mathbb{Q}(\theta).$$

Let $X : \Omega \rightarrow \mathcal{X}$ be a random element to measurable space $(\mathcal{X}, \mathcal{H})$ and suppose that (θ, ω) is sampled from the joint distribution \mathbb{P} and $X(\omega) = x$ is observed. The posterior should be a measure on (Θ, \mathcal{G}) that depends on the observed data. In other words, it should be a probability kernel from $(\mathcal{X}, \mathcal{H})$ to (Θ, \mathcal{G}) . Without much thought we might try and apply Bayes' law (Eq. 2.1) to claim that the posterior distribution having observed $X(\omega) = x$ should be a measure on (Θ, \mathcal{F}) given by

$$\mathbb{Q}(A \mid X = x) = \mathbb{P}(\theta \in A \mid X = x) = \frac{\mathbb{P}(X = x \mid \theta \in A)\mathbb{P}(\theta \in A)}{\mathbb{P}(X = x)}.$$

The problem is that $\mathbb{P}(X = x)$ can have measure zero and then $\mathbb{P}(\theta \in A \mid X = x)$

is not defined. This is not an esoteric problem. When $\Theta = \Omega = \mathbb{R}$ with the usual Borel σ -algebras and $\mathbb{P}_\theta = \mathcal{N}(\theta, 1)$ is the Gaussian family and $X(\omega) = \omega$, then $\mathbb{P}(X = x) = 0$ for all x . Having read Chapter 2, the next attempt might be to define $\mathbb{Q}(A | X)$ as a $\sigma(X)$ -measurable random variable defined using the conditional expectation.

$$\mathbb{Q}(A | X) = \mathbb{P}(\theta \in A | X) = \mathbb{E}[\mathbb{I}_A(\theta(\omega)) | X].$$

Recall that $\mathbb{E}[\mathbb{I}_A(\theta) | X]$ is a $\sigma(X)$ -measurable random variable that is uniquely defined except for a set of measure zero. The nonuniqueness means that $\mathbb{Q}(A | X)$ is actually a version of $\mathbb{P}(\theta \in A | X)$ and which version should really be specified. For most applications of probability theory the choice of conditional expectation does not matter, but this is not true here. Perhaps a more annoying issue than nonuniqueness is that $\mathbb{Q}(\cdot | X)$ as defined above need not be a measure and the classical theorems on the existence of conditional expectations do not guarantee such a choice exists. Provided that (Θ, Ω) is not too big, however, one can guarantee the existence of a conditional expectation satisfying the conditions of a Markov kernel.

THEOREM 34.2 *If (Θ, \mathcal{G}) is a Borel space, then there exists a probability kernel $\mathbb{Q} : \mathcal{X} \times \mathcal{G} \rightarrow [0, 1]$ such that $\mathbb{Q}(A | X) = \mathbb{P}(\theta \in A | X)$ almost surely for all $A \in \mathcal{G}$ and $\mathbb{Q}(\cdot | X)$ is unique $\mathbb{P} \circ X^{-1}$ -a.s.*



The notation for a probability kernel differs from the notation introduced in Chapter 3 because here we want to emphasize the fact that $\mathbb{Q}(A | X)$ is derived by conditioning. There may also be some confusion about the usage of $\mathbb{P}(\theta \in A | X) = \mathbb{E}[\mathbb{I}_A(\theta) | X]$, which by definition is a $\sigma(X)$ -measurable random variable on $\Theta \times \Omega$. Because it is $\sigma(X)$ -measurable, however, by Lemma 2.1 there exists a \mathcal{H}/\mathcal{G} -measurable function $f : \mathcal{X} \rightarrow \Theta$ such that $\mathbb{P}(\theta \in A | X) = f \circ X$ so that really $\mathbb{P}(\theta \in A | X)$ can be viewed as a function from \mathcal{X} . The theorem above shows that it can be chosen so that $\mathbb{P}(\theta \in \cdot | X)$ is also a measure.

Theorem 34.2 shows the posterior exists, but does not suggest a useful way of finding it. In many practical situations the posterior can be calculated using densities. Given $\theta \in \Theta$ let $p_\theta(x)$ be the Radon-Nikodym derivative of $(\mathbb{P}_\theta)_X$ with respect to some measure μ and let $q(\theta)$ be the Radon-Nikodym derivative of \mathbb{Q} with respect to another measure ν . Provided all terms are appropriately measurable and nonzero, then

$$q(\theta | X) = \frac{p_\theta(X)q(\theta)}{\int_\Theta p_\theta(X)q(\theta)d\nu(\theta)}$$

is the Radon-Nikodym derivative of $\mathbb{Q}(\cdot | X)$ with respect to ν . In other words, for any $A \in \mathcal{G}$ it holds that $\mathbb{Q}(A | X) = \int_A q(\theta | X)d\nu(\theta)$. This corresponds to the usual manipulation of densities when μ and ν are the Lebesgue measures.

EXAMPLE 34.1 To emphasize the nonuniqueness of the posterior, let $\Theta = [0, 1]$ and \mathbb{Q} be the uniform measure on Θ and $\mathbb{P}_\theta = \delta_\theta$ be the Dirac measure on $[0, 1]$ at θ . The following posterior satisfies the conditions of Theorem 34.2 for any measure μ on $([0, 1], \mathcal{B}(\mathbb{R}))$ and countable $C \subset [0, 1]$.

$$\mathbb{Q}(A | X = x) = \begin{cases} \delta_x(A) & \text{if } x \notin C \\ \mu(A) & \text{if } x \in C. \end{cases}$$

A true Bayesian is probably unconcerned. If θ is sampled from the prior \mathbb{Q} , then the event $\{X \in C\}$ has measure zero and there is little cause to worry about events that happen with probability zero. But for a frequentist using Bayesian techniques for inference this actually matters. If θ is not sampled from \mathbb{Q} , then nothing prevents the situation that $\theta \in C$ and the nonuniqueness of the posterior is an issue. Probability theory does not provide a way around this issue.



When using Bayesian techniques for inference in a frequentist setting one should be careful to specify the version of the posterior being used. This is important because in the frequentist viewpoint θ is not part of the probability space and results are proven for \mathbb{P}_θ . By contrast, the all-in Bayesian includes θ in the probability space and need not worry about events with negligible prior probability.

34.3 Conjugate priors and the exponential family (\dagger)

One of the strengths of Bayesian methods is the ability to incorporate prior knowledge into the algorithm in a natural way via the prior. This advantage is belied a little by the competing necessity of choosing a prior for which the posterior can be efficiently computed. The ease of computing the posterior depends on the interplay between the prior and the model. Given the importance of computation, it is hardly surprising that researchers have worked hard to find models and priors that behave well together. A prior and model are called **conjugate** if the posterior has the same parametric form as the prior.

Gaussian model/Gaussian prior

Suppose that $(\Theta, \mathcal{G}) = (\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $X : \Omega \rightarrow \omega$ is the identity and \mathbb{P}_θ is Gaussian with mean θ and known **signal variance** σ_S^2 . If the prior \mathbb{Q} is Gaussian with mean μ_P and **prior variance** σ_P^2 , then the posterior distribution having observed $X = x$ is

$$\mathbb{Q}(\cdot | X = x) = \mathcal{N}\left(\frac{\mu_P/\sigma_P^2 + x/\sigma_S^2}{1/\sigma_P^2 + 1/\sigma_S^2}, \left(\frac{1}{\sigma_S^2} + \frac{1}{\sigma_P^2}\right)^{-1}\right).$$

We leave the proof of this fact to the reader (Exercise 34.2). The limiting regimes as the prior/signal variance tend to zero or infinity are quite illuminating. For

example, as $\sigma_P^2 \rightarrow 0$ the posterior tends to a Gaussian $\mathcal{N}(\mu_P, \sigma_P^2)$, which is *equal* to the prior and indicates that no learning occurs. This is consistent with intuition: If the prior variance is zero, then the statistician is already certain of the mean and no amount of data can change their belief. On the other hand, as σ_P^2 tends to infinity we see the mean of the posterior has no dependence on the prior mean, which means that all prior knowledge is washed away with just one sample. We encourage you to examine what happens when $\sigma_S^2 \rightarrow \{0, \infty\}$.

Notice how the model has fixed σ_S^2 , suggesting that the model variance is known. We made this kind of assumption very often in the book so far, but the Bayesian can incorporate their uncertainty over the variance. In this case the model parameters are $\Theta = \mathbb{R} \times [0, \infty)$ and $\mathbb{P}_\Theta = \mathcal{N}(\theta_1, \theta_2)$. But is there a conjugate prior in this case? Already things are getting complicated, so we will simply let you know that the family of Gaussian-inverse-gamma distributions is conjugate.

Bernoulli model/beta prior

Suppose that $\Theta = [0, 1]$ and $\mathbb{P}_\theta = \mathcal{B}(\theta)$ is Bernoulli with parameter θ . In this case it turns out that the family of beta distributions is conjugate, which for parameters $\theta = (\alpha, \beta) \in (0, \infty)^2$ is given in terms of its probability density function with respect to the Lebesgue measure:

$$p_{\alpha, \beta}(x) = x^{\alpha-1}(1-x)^{\beta-1} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)},$$

where $\Gamma(x)$ is the Gamma function. Then the posterior having observed $X \in \{0, 1\}$ is also a beta distribution with parameters $(\alpha + X, \beta + 1 - X)$.

Exponential families

Both the Gaussian and Bernoulli families are examples of a more general concept. Let μ be a measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ and $T, \eta : \mathbb{R} \rightarrow \mathbb{R}$ where T is called the **sufficient statistic** and define measure \mathbb{P}_θ on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ in terms of its Nikodym derivatives with respect to μ .

$$\frac{d\mathbb{P}_\theta}{d\mu}(x) = \exp(\eta(\theta)T(x) - A(\theta)),$$

where $A(\theta) = \log \int_{\mathbb{R}} \exp(\eta(\theta)T(x)) d\mu(x)$ is the **log partition** function. Let $\Theta = \text{dom}(A) = \{\theta : A(\theta) < \infty\}$ be the domain of A and for $\theta \in A$ define measure P_θ on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ by

$$P_\theta(A) = \int_A \frac{dP_\theta}{d\mu}(x) d\mu(x).$$

The collection $\{P_\theta : \theta \in \Theta\}$ is called a **single parameter exponential family**.

EXAMPLE 34.2 Let $\sigma^2 > 0$ and $\mu = \mathcal{N}(0, \sigma^2)$ and $\eta(\theta) = \frac{\theta}{\sigma}$ and $T(x) = \frac{x}{\sigma}$. An easy calculation shows that $A(\theta) = \theta^2/(2\sigma^2)$, which has domain $\Theta = \mathbb{R}$ and $P_\theta = \mathcal{N}(\theta, \sigma^2)$.

EXAMPLE 34.3 Let $\mu = \delta_0 + \delta_1$ be the sum of Dirac measures and $T(x) = x$ and $\eta(\theta) = \theta$. Then $A(\theta) = \log(1 + \exp(\theta))$ and $\Theta = \mathbb{R}$ and $P_\theta = \mathcal{B}(\sigma(\theta))$ where $\sigma(\theta) = \exp(\theta)/(1 + \exp(\theta))$ is the sigmoid function.

EXAMPLE 34.4 The same family can be parameterized in many different ways. Let $\mu = \delta_0 + \delta_1$ and $T(x) = x$ and $\eta(\theta) = \log(\theta/(1 - \theta))$. Then $A(\theta) = -\log(1 - \theta)$ and $\Theta = (0, 1)$ and $P_\theta = \mathcal{B}(\theta)$.

Exponential families have many nice properties. Of most interest to us here is the existence of conjugate priors. Suppose that $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is a single parameter exponential family determined by functions η and T with $T(x) = x$ assumed to be the identity. Let $x_0, n_0 \in \mathbb{R}$ and define prior measure \mathbb{Q} on $(\Theta, \mathcal{B}(\Theta))$ in terms of its density $q = d\mathbb{Q}/d\lambda$ with λ the Lebesgue measure.

$$q(\theta) = \frac{\exp(n_0 x_0 \eta(\theta) - n_0 A(\theta))}{\int_{\Theta} \exp(n_0 x_0 \eta(\theta) - n_0 A(\theta)) d\theta},$$

where we assume the existence and strict positivity of the integral in the denominator. Suppose we observe $X = x$, then the posterior has density with respect to the Lebesgue measure given by

$$q(\theta | x) = \frac{\exp(\eta(\theta)(x + n_0 x_0) - (1 + n_0)A(\theta))}{\int_{\Theta} \exp(\eta(\theta)(x + n_0 x_0) - (1 + n_0)A(\theta)) d\lambda(\theta)}.$$

To see how this recovers existing results consider the Bernoulli case of Example 34.4,

$$\exp(n_0 x_0 \eta(\theta) - n_0 A(\theta)) = \left(\frac{\theta}{1 - \theta} \right)^{n_0 x_0} (1 - \theta)^{n_0} = \theta^{n_0 x_0} (1 - \theta)^{n_0(1 - x_0)},$$

which is proportional to a beta distribution with $\alpha = 1 + n_0 x_0$ and $\beta = 1 + n_0(1 - x_0)$.



There are important parametric families with conjugate priors that are not exponential families. One example is the uniform family $\{\mathcal{U}(a, b) : a < b\}$, which is conjugate to the Pareto family.

34.4 Bayesian learning and bandits

Adapting the tools of the previous two sections to bandits is straightforward. Let \mathcal{E} be a set of K -armed stochastic bandits and \mathcal{G} be a σ -algebra on \mathcal{E} and \mathbb{Q} be a prior measure on $(\mathcal{E}, \mathcal{G})$. Given a bandit $\nu \in \mathcal{E}$ let \mathbb{P}_ν be the product measure on $(\mathbb{R}^K, \mathcal{B}(\mathbb{R}^K))$ that corresponds to the reward distributions. We assume that \mathbb{P}_ν is a Markov kernel from $(\mathcal{E}, \mathcal{G})$ to $(\mathbb{R}^K, \mathcal{B}(\mathbb{R}^K))$. Fix a policy $\pi = (\pi_1, \dots, \pi_n)$ and let

$$\Omega = \{a_1, x_1, \dots, a_n, x_n : a_t \in [K] \text{ and } x_t \in \mathbb{R}^K\} \quad \text{and} \quad \mathcal{F} = (\rho \otimes \mathcal{B}(\mathbb{R}))^n. \tag{34.2}$$

Then define joint probability space $(\mathcal{E} \times \Omega, \mathcal{G} \otimes \mathcal{F}, \mathbb{P})$ where $\mathbb{P} = \mathbb{Q} \otimes \pi_1 \otimes P_\nu \otimes \dots \otimes \pi_n \otimes P_\nu$. The coordinate projections are

$$\begin{aligned}\nu((\nu, a_1, x_1, \dots, a_n, x_n)) &= \nu \quad \text{and} \\ A_t((\nu, a_1, x_1, \dots, a_n, x_n)) &= a_t \quad \text{and} \\ X_t((\nu, a_1, x_1, \dots, a_n, x_n)) &= x_t,\end{aligned}\tag{34.3}$$

which means that $\nu \in \mathcal{E}$, $A_t \in [K]$ and $X_t \in \mathbb{R}^K$ are random elements. Then let $\mathcal{F}_t = \sigma(A_1, X_{1A_1}, \dots, A_t, X_{tA_t})$ be the σ -algebra generated by the observations of the learner after t rounds. The posterior after t rounds is a probability kernel from (Ω, \mathcal{F}_t) to $(\mathcal{E}, \mathcal{G})$ denoted by $\mathbb{Q}_t(\cdot) = \mathbb{Q}(\cdot | A_1, X_1, \dots, A_t, X_t)$ that for all $B \in \mathcal{G}$ satisfies

$$\mathbb{Q}_t(B) = \mathbb{E}[\mathbb{I}_B(\nu) | A_1, X_1, \dots, A_t, X_t] \quad \text{a.s.}$$

Theorem 34.2 guarantees the existence of the posterior as long as $(\mathcal{E}, \mathcal{G})$ is a Borel space, but the abstract definition is not very useful for explicit calculations. By making mild assumptions the posterior can be written in terms of densities. Assume there exists a σ -finite measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $P \ll \lambda$ for all reward distributions P used by the bandits in \mathcal{E} . Recall from Chapter 15 that the Radon-Nikodym derivative of $\mathbb{P}_{\nu\pi}$ with respect to $(\rho \times \lambda)^n$ is

$$p_{\nu\pi}(a_1, x_1, \dots, a_n, x_n) = \prod_{t=1}^n \pi_t(a_t | a_1, x_1, \dots, a_{t-1}, x_{t-1}) p_{\nu a_t}(x_t), \tag{34.4}$$

where $p_{\nu a}$ is the density with respect to λ of the reward distribution for the a th arm of ν . Then the posterior after t rounds is given by

$$\begin{aligned}\mathbb{Q}(B | a_1, x_1, \dots, a_t, x_t) &= \frac{\int_B p_{\nu\pi}(a_1, x_1, \dots, a_t, x_t) d\mathbb{Q}(\nu)}{\int_{\mathcal{E}} p_{\nu\pi}(a_1, x_1, \dots, a_t, x_t) d\mathbb{Q}(\nu)} \\ &= \frac{\int_B \prod_{s=1}^t p_{\nu a_s}(x_s) d\mathbb{Q}(\nu)}{\int_{\mathcal{E}} \prod_{s=1}^t p_{\nu a_s}(x_s) d\mathbb{Q}(\nu)},\end{aligned}\tag{34.5}$$

where the second equality follows from Eq. (34.4). Of course we are assuming here that all quantities are well defined. In particular, the integral in the denominator must be positive almost surely and $p_{\nu a}(x)$ should be measurable as a function of ν for all x .

EXAMPLE 34.5 Let $\mathcal{E} = \mathcal{E}_{\mathcal{B}}^K$ be the set of all Bernoulli bandits with K arms. Bandits in \mathcal{E} are characterized by their mean vectors in $[0, 1]^K$ so it suffices to choose our prior on $[0, 1]^K$ with the Lebesgue σ -algebra. A natural prior is a product of Beta priors with parameters $\alpha, \beta > 0$ defined in terms of its density with respect to the Lebesgue measure λ by

$$q(\theta) \propto \prod_{i=1}^K \theta_i^{\alpha-1} (1 - \theta_i)^{\beta-1}.$$

Recall that $T_i(t) = \sum_{s=1}^t \mathbb{I}\{A_t = i\}$ and let $S_i(t) = \sum_{s=1}^t \mathbb{I}\{A_t = i\} X_t$. The

posterior is also given in terms of its density with respect to the Lebesgue measure.

$$q(\theta | A_1, X_1, \dots, A_t, X_t) \propto \prod_{i=1}^K \theta_i^{\alpha+S_i(t)-1} (1-\theta_i)^{\beta+T_i(t)-S_i(t)-1}.$$

This means the posterior is also the product of beta distributions, each updated according to the observations from the relevant arm.

34.5 One-armed bandits

We return to the one-armed bandit problem that appeared in various exercises in earlier chapters. In each round t the learner chooses action $A_t \in \{1, 2\}$. The reward when choosing the first action is X_t where X_1, \dots, X_n is a sequence of independent and identically distributed random variables with unknown distribution and mean $\mu \in \mathbb{R}$. The reward when choosing the second action is a deterministic known value $\mu_o \in \mathbb{R}$. In Exercise 4.9 we defined a retirement policy for one-armed bandits as a policy that chooses $A_t = 1$ until some random time and subsequently $A_t = 2$. There you showed that provided the horizon n is known in advance, then there is no reason to consider policies of any other kind. So the problem reduces to finding the ‘best’ retirement policy.

In order to be rigorous we need a probability space to hold all these random variables. Let ν be a measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and \mathbb{P}_ν be the product measure ν^n on $(\Omega, \mathcal{F}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. The reward of the first arm after round t is random variable $X_t : \mathbb{R}^n \rightarrow \mathbb{R}$ defined by $X_t(\omega) = \omega_t$. Let $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ so that a retirement policy is a stopping time $0 \leq \tau \leq n$ with respect to filtration $\mathbb{F} = (\mathcal{F}_t)_{t=0}^n$. The frequentist regret of the policy induced by \mathbb{F} -stopping time τ is

$$\begin{aligned} R_n(\tau, \nu) &= n \max \{\mu, \mu_o\} - \mathbb{E} \left[\sum_{t=1}^{\tau} X_t + \sum_{t=\tau+1}^n \mu_o \right] \\ &= \mathbb{E} \left[\sum_{t=1}^{\tau} \max\{0, \mu_o - \mu\} + \sum_{t=\tau+1}^n \max\{\mu - \mu_o, 0\} \right]. \end{aligned}$$

The regret is minimized by deterministic stopping time $\tau = n$ if $\mu > \mu_o$ and $\tau = 0$ otherwise. As usual the problem is that μ is unknown.

Frequentist regret and policy

If we assume that ν is 1-subgaussian, then the techniques of Part II can be applied to derive a stopping time such that

$$R_n(\tau, \nu) \leq \begin{cases} \Delta & \text{if } \Delta \geq 0 \\ \min\{\Delta + C \log(n)/\Delta, n\Delta\} & \text{otherwise,} \end{cases} \quad (34.6)$$

where $C > 0$ is a universal constant and $\Delta = |\mu - \mu_\circ|$. An example stopping time for which this holds is

$$\tau = n \wedge \min \left\{ t \in [n] : \hat{\mu}_t + \sqrt{\frac{2 \log(n^2)}{t}} \leq \mu_\circ \right\}, \quad (34.7)$$

where $\hat{\mu}_t = \frac{1}{t} \sum_{s=1}^t X_s$. We leave it to the reader to establish that Eq. (34.6) indeed holds for the retirement policy using the above stopping time (Exercise 34.7).

Bayesian regret and policy

Moving on to the Bayesian framework, we now suppose that $\nu = \nu_\theta$ where $\theta \in \Theta$ and $\{\mathbb{P}_{\nu_\theta} : \theta \in \Theta\}$ is a probability kernel from Borel space (Θ, \mathcal{G}) to (Ω, \mathcal{F}) . Then let \mathbb{Q} be a prior measure on (Θ, \mathcal{G}) and \mathbb{P} be the measure on $(\Theta \times \Omega, \mathcal{G} \otimes \mathcal{F})$ given by

$$\mathbb{P}(\theta \in A, \omega \in B) = \int_A \mathbb{P}_{\nu_\theta}(B) d\mathbb{Q}(\theta).$$

Unless otherwise specified, expectations for the remainder of the section are with respect to \mathbb{P} .



In the model used in the frequentist setting the variables X_1, X_2, \dots, X_n are independent and identically distributed with respect to \mathbb{P}_ν . Having incorporated θ into the probability space this is not true anymore. Up to the usual ‘almost surely’ exceptions they are conditionally independent and identically distributed given θ .

The posterior after t observations is a probability kernel \mathbb{Q}_t from $(\mathbb{R}^t, \mathfrak{L}(\mathbb{R}^t))$ to (Θ, \mathcal{G}) such that $\mathbb{Q}_t(A) = \mathbb{E}[\mathbb{I}_A(\theta) \mid \mathcal{F}_t]$ almost surely. We abbreviate $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. The Bayesian regret of the retirement policy determined by τ is

$$\text{BR}_n(\tau, \mathbb{Q}) = \mathbb{E}[R_n(\tau, \nu_\theta)].$$

The Bayesian optimal policy (if it exists) is a retirement policy that minimizes this quantity,

$$\tau^* \in \operatorname{argmin}_\tau \text{BR}_n(\tau, \mathbb{Q}) = \operatorname{argmax}_\tau \mathbb{E} \left[\sum_{t=1}^\tau X_t + (n - \tau)\mu_\circ \right].$$

The key idea to finding τ^* is to rewrite the optimization problem in terms of an **optimal stopping problem**. Let $U_t = \sum_{s=1}^t X_s + (n - t)\mu_\circ$ be the cumulative reward received by the learner when $\tau = t$. Then

$$\tau^* \in \operatorname{argmax}_\tau \mathbb{E}[U_\tau]. \quad (34.8)$$

Because U_t is \mathcal{F}_t -measurable, this problem is called a **standard optimal stopping problem**. For standard problems in discrete time with a finite horizon the solution can be found using **backwards induction**. Intuitively, having

observed X_1, \dots, X_t the optimal policy will retire if U_t is larger than the expected return from the optimal stopping policy that stops after t . This suggests defining things backwards from n by

$$V_n = U_n \quad \text{and} \quad V_t = \max\{U_t, \mathbb{E}_t[V_{t+1}]\} \quad \text{for } t < n.$$

The process $(V_t)_t$ is called the **Snell envelope** and the optimal stopping time stops at the earliest time such that $U_t \geq V_t$. This intuitive fact is captured in the following theorem.

THEOREM 34.3 *Assuming that $\sup_\tau \mathbb{E}[|U_\tau|] < \infty$, then $\tau^* = \min\{t : U_t \geq V_t\}$ satisfies Eq. (34.8).*

The optimal policy in Theorem 34.3 only depends on the ordering of U_t and V_t , which by subtracting the cumulative observed reward allows us to rewrite the optimal stopping time in a more convenient form. Define $W_n = 0$ and for $t < n$ let $W_t = V_t - \sum_{s=1}^t X_s$, which satisfies

$$\begin{aligned} W_t &= \max \left((n-t)\mu_o, \mathbb{E}_t[V_{t+1}] - \sum_{s=1}^t X_s \right) \\ &= \max ((n-t)\mu_o, \mathbb{E}_t[W_{t+1}] + \mathbb{E}_t[X_{t+1}]). \end{aligned} \quad (34.9)$$

Then the optimal policy is

$$\begin{aligned} \tau^* &= \min\{t : U_t \geq V_t\} = \min \left\{ t : U_t - \sum_{s=1}^t X_s \geq V_t - \sum_{s=1}^t X_s \right\} \\ &= \min \{t : (n-t)\mu_o \geq \mathbb{E}_t[W_{t+1}] + \mathbb{E}_t[X_{t+1}]\}. \end{aligned}$$

Theorem 34.3 and the above display characterize the optimal stopping rule in a straightforward way. The difficulty is that $\mathbb{E}_t[W_{t+1}]$ is usually a complicated object. We now give two examples where $\mathbb{E}_t[W_{t+1}]$ has a simple representation that means computing the optimal stopping rule is practical.

Bernoulli rewards

Let $\Theta = [0, 1]$ and \mathcal{F} be the standard Borel σ -algebra and $\nu_\theta = \mathcal{B}(\theta)$ be Bernoulli with bias θ . In the previous section we showed that the beta prior and Bernoulli family are conjugates so we will choose the prior to be $\mathbb{Q} = \text{Beta}(\alpha, \beta)$ for some $\alpha, \beta > 0$. A calculation shows that

$$\mathbb{E}[X_{t+1} | \mathcal{F}_t] = \frac{\alpha + S_t}{\alpha + \beta + t} = p_t(S_t),$$

where $p_t(s) = (\alpha + s)/(\alpha + \beta + t)$. This greatly simplifies matters because W_t can be written as a function of just $S_t \in \{0, 1, \dots, t\}$.

$$\begin{aligned} W_t(s) &= \max ((n-t)\mu_o, \mathbb{E}[W_{t+1} | S_t = s] + \mathbb{E}_t[X_{t+1} | S_t = s]) \\ &= \max ((n-t)\mu_o, p_t(s)W_{t+1}(s+1) + (1 - p_t(s))W_{t+1}(s) + p_t(s)). \end{aligned}$$

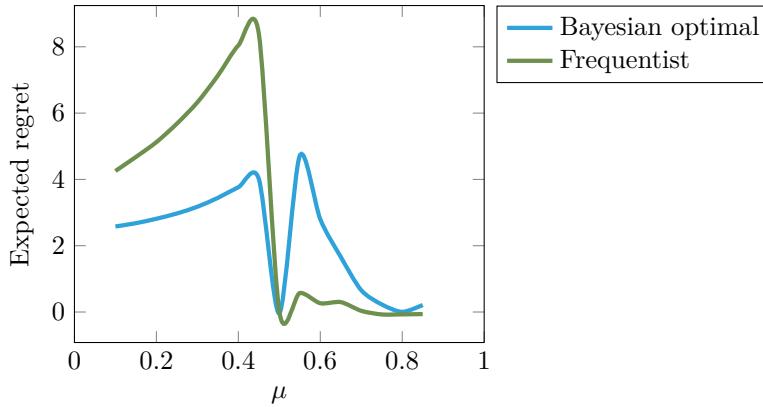
So the optimal policy can be computed by evaluating $W_t(s)$ for all $s \in \{0, \dots, t\}$ starting with $t = n - 1$, then $n - 2$ and so-on until $t = 0$. The total computation for this backwards induction is $O(n^2)$ and the output is a policy that can be implemented over all n rounds. In contrast, the stopping rule proposed in Eq. (34.7) requires only $O(n)$ computations, so the overhead is quite severe. The improvement is also not insignificant as illustrated by the following experiment.



The horizon is set to $n = 500$ and $\mu_0 = 1/2$. The stopping times we compare are the Bayesian optimal policy with a Beta(1, 1) prior and the ‘frequentist’ stopping time given by

$$\tau_D = n \wedge \min \left\{ t \geq 1 : \hat{\mu}_t < \mu_0 \text{ and } d(\hat{\mu}_t, \mu_0) \geq \frac{\log(n/t)}{t} \right\}, \quad (34.10)$$

where $d(p, q) = D(\mathcal{B}(p), \mathcal{B}(q))$ is the relative entropy between Bernoulli distributions with parameters p and q respectively. The plot below shows the expected regret for different values of μ . As you can see, the results are not a clear win in favour of the Bayesian optimal policy. The asymmetric behaviour of the frequentist policy is explained by the conservatism of the confidence interval in Eq. (34.10), which makes it stop consistently later than its Bayesian counterpart. In a sense this is an advantage of the Bayesian approach, where the prior encodes the objective and the policy automatically optimises the criteria. Because the Beta(1, 1) prior is symmetric about 1/2 it should not surprise us that the regret is approximately symmetric.



Gaussian rewards

The Gaussian case is more delicate because W_t does not have a discrete representation. To make things concrete assume that $\nu_\theta = \mathcal{N}(\theta, 1)$ is Gaussian with unit variance and the prior \mathbb{Q} on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is also Gaussian with mean $\mu_P \in \mathbb{R}$ and variance $\sigma_P^2 > 0$. By the results in Section 34.3 the posterior \mathbb{Q}_t is

Gaussian with mean μ_t and variance σ_t^2 given by

$$\mu_t = \frac{\frac{\mu_P}{\sigma_P^2} + \sum_{s=1}^t X_s}{1 + \sigma_P^{-2}} \quad \text{and} \quad \sigma_t^2 = \left(t + \frac{1}{\sigma_P^2} \right)^{-1}.$$

Notice that σ_t^2 is independent of the observations so the posterior is determined entirely by its mean. Thus we can view W_t as a function from the posterior mean $\mu_t \in \mathbb{R}$ to \mathbb{R} , which by Eq. (34.9) is given by

$$W_t(\mu) = \max \left((n-t)\mu_o, \mu + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left(-\frac{x^2}{2\sigma_t^2} \right) W_{t+1}(\mu+x) dx \right).$$

In general the integral on the right-hand side does not have a closed form solution, which forces the use of approximate methods. Fortunately W_t is a well behaved function and can be efficiently approximated.

LEMMA 34.1 *The following hold:*

- (a) $W_t(\mu)$ is monotone increasing in μ .
- (b) $W_t(\mu)$ is convex.
- (c) $\lim_{\mu \rightarrow \infty} W_t(\mu)/\mu = n-t$ and $\lim_{\mu \rightarrow -\infty} W_t(\mu) = (n-t)\mu_o$.

Proof The first two follow from the calculus of monotone/convex functions while the third we leave as an exercise to the reader. \square

There are many ways to approximate a function, but the important point here is we want an approximation of W_t such that the integral in the recursive definition can be computed efficiently. Given the properties in Lemma 34.1 a natural choice is to approximate W_t using piecewise quadratic functions. Let $\tilde{W}_{n+1}(\mu) = 0$ and

$$\bar{W}_t(\mu) = \max \left\{ (n-t)\mu_o, \mu + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left(-\frac{x^2}{2\sigma_t^2} \right) \tilde{W}_{t+1}(\mu+x) dx \right\}.$$

Then let $-\infty < x_1 \leq x_2 \leq \dots \leq x_N < \infty$ and for $\mu \in [x_i, x_{i+1}]$ define $\tilde{W}_t(\mu) = a_i \mu^2 + b_i \mu + c_i$ to be the unique quadratic approximation of $\bar{W}_t(\mu)$ such that

$$\begin{aligned} \tilde{W}_t(x_i) &= \bar{W}_t(x_i) \\ \tilde{W}_t(x_{i+1}) &= \bar{W}_t(x_{i+1}) \\ \tilde{W}_t((x_i + x_{i+1})/2) &= \bar{W}_t((x_i + x_{i+1})/2). \end{aligned}$$

For $\mu < x_1$ we approximate $W_t(\mu) = (n-t)\mu_o$ and for $\mu > x_N$ the linear approximation $\tilde{W}_t(\mu) = (n-t)\mu$ is reasonable by Lemma 34.1. The computation time for calculating the coefficients a_i, b_i, c_i for all t and $i \in [N]$ is $O(Nn)$.

34.6 Gittins index

Generalizing the analysis in the previous section to multiple actions is mathematically straightforward, but computationally intractable. The computational complexity of backwards induction increases exponentially with the number arms, which even for two actions makes this approach quite impractical.

An **index policy** is a policy that in each round computes a real-valued **index** for each arm and plays the arm with the largest index. Furthermore, the index of an arm should only depend on statistics collected for that arm (and perhaps the time horizon). For example, most variants of the upper confidence bound algorithm introduced in Part II are index policies. Sadly, however, the Bayesian optimal policy for finite-horizon bandits is not usually an index policy. John Gittins proved that if one is prepared to modify the objective to a special kind of infinite horizon problem, then the Bayesian optimal policy becomes an index policy.

A discounted retirement game

We start by describing the discounted setting with one action and then generalize to multiple actions. Let $(\mathcal{S}, \mathcal{G})$ be a measurable space called the **state space**. Then let μ be a Markov kernel from $(\mathcal{S}, \mathcal{G})$ to itself and $(\Omega, \mathcal{F}, \mathbb{P}_s)$ be a probability space and S_1, S_2, \dots be a sequence of \mathcal{F}/\mathcal{G} -measurable random elements such that $\mathbb{P}_s(S_1 = s) = 1$ and $\mathbb{P}_s(S_{t+1} \in \cdot | S_t) = \mu(S_t, \cdot)$ almost surely. Finally let $r : \mathcal{S} \rightarrow [0, 1]$ be a known measurable function and $\gamma \in \mathbb{R}$. In each round t the learner observes the state S_t and chooses one of two options: (a) to retire, which ends the game. Or (b) pay a fixed cost of γ to receive a reward of $r(S_t)$ and continue for another round. The policy of a learner in this game corresponds to choosing a stopping time τ with respect to the filtration $\mathbb{F} = (\mathcal{F}_t)_t$ with $\mathcal{F}_t = \sigma(S_1, \dots, S_t)$, where $\tau = t$ means that the learner retires after observing S_t at the start of round t . The value of a retirement policy τ is given by

$$V^\tau(s; \gamma) = \mathbb{E}_s \left[\sum_{t=1}^{\tau-1} \alpha^{t-1} (r(S_t) - \gamma) \right],$$

where $\alpha \in (0, 1)$ is the **discount factor** and \mathbb{E}_s is the expectation with respect to \mathbb{P}_s . This definition of the value function means a learner is encouraged to obtain large rewards earlier rather than later and is one distinction between this model and the finite-horizon model studied for most of this book. A brief discussion of discounting is left for the notes.

If $\tau = 1$ almost surely, then learner retires immediately without receiving any reward or paying any cost and $V^\tau(s; \gamma) = 0$. The **Gittins index** or **fair charge** of a state s is the largest value of γ for which the learner is indifferent between retiring immediately and playing for at least one round:

$$G^*(s) = \sup \left\{ \gamma \in \mathbb{R} : \sup_{\tau > 1} V^\tau(s; \gamma) \geq 0 \right\}, \quad (34.11)$$

where the inner supremum is taken over \mathbb{F} -stopping times τ with $\tau > 1$ almost surely. Straightforward manipulation (Exercise 34.4) shows that

$$G^*(s) = \sup_{\tau > 1} \frac{\mathbb{E}_s \left[\sum_{t=1}^{\tau-1} \alpha^t r(S_t) \right]}{\mathbb{E}_s \left[\sum_{t=1}^{\tau-1} \alpha^t \right]}, \quad (34.12)$$

It is not immediately clear that a stopping time attaining the supremum in the definition exists. The following lemma shows that it does and gives an explicit form. The proof of this result is left as a technical challenge for the reader (Exercise 34.5).

LEMMA 34.2 *For each $s \in \mathcal{S}$ the following stopping times both attain the supremum in Eq. (34.12).*

- (a) $\tau = \min\{t > 1 : G^*(S_t) < G^*(s)\}$.
- (b) $\tau = \min\{t > 1 : G^*(S_t) \leq G^*(s)\}$.

The result is relatively intuitive. The Gittins index represents the price the learner should be willing to pay for the privilege of continuing to play. The optimal policy continues to play as long as the actual value of the game is not smaller than this price with an indifference region when the price is exactly equal to the value.

Discounted bandits and the index theorem

The generalization of the discounted retirement game to multiple arms is quite straightforward. There are now K independent Markov chains on the same state-space and in each round the learner first observes the state of all chains and chooses an action $A_t \in [K]$. The learner receives a reward from the corresponding chain, which then evolves randomly to a new state sampled from the probability kernel. The states for unplayed arms do not change and we assume that all chains evolve according to the same Markov kernel. The protocol is given in Fig. 34.1.

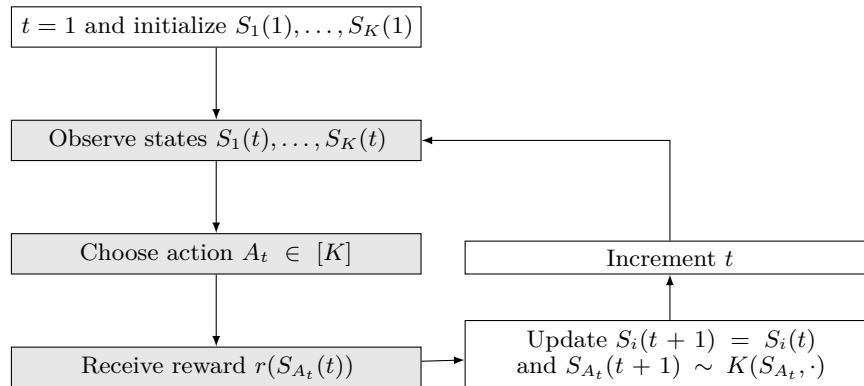


Figure 34.1 Interaction protocol for discounted bandits.



The assumption that the Markov chains evolve on the same state-space with the same transition kernel is non-restrictive since the state-space can always be taken to be the union of K state-spaces and the transition kernel defined with K disconnected components.

Given a discount parameter $\alpha \in (0, 1)$, the value of policy π is

$$V^\pi = \mathbb{E} \left[\sum_{t=1}^{\infty} \alpha^t r(S_{A_t}(t)) \right].$$

EXAMPLE 34.6 To see the relation to Bayesian bandits with discounted rewards consider the following setup. Let $\mathcal{S} = [0, \infty) \times [0, \infty)$ and $\mathcal{G} = \mathfrak{B}(\mathcal{S})$. Then let the initial state of each Markov chain be $S_i(1) = (1, 1)$ and define probability kernel μ from $(\mathcal{S}, \mathcal{G})$ to itself by

$$\mu((x, y), A) = \frac{x}{x+y} \delta_{(x+1, y)}(A) + \frac{y}{x+y} \delta_{(x, y+1)}(A).$$

The reward function is $r(x, y) = x/(x+y)$. The reader should check that this corresponds to a Bernoulli bandit with Beta(1,1) prior on the mean reward of each arm.

One of the most celebrated theorems in the study of bandits is that the optimal policy for this problem is to choose in each round the Markov chain with the largest Gittins index.

THEOREM 34.4 *Let π^* be the policy choosing $A_t = \operatorname{argmax}_i G^*(S_i(t))$. Then $V^{\pi^*} = \sup_\pi V^\pi$ where the supremum is taken over all policies.*

The remainder of the section is devoted to proving Theorem 34.4. The choice of actions produces an interleaving of the rewards generated by each Markov chain and it will be useful to have a notation for these interleavings. For each $i \in [K]$ let $g_i = (g_{it})_{t=1}^\infty$ be a real-valued sequence and $g = (g_1, \dots, g_K)$ be the tuple of these sequences. Given an infinite sequence $(a_t)_{t=1}^\infty$ with $a_t \in [K]$ define the interleaving sequence $I_1(g, a), I_2(g, a), \dots$ by

$$I_t(g, a) = g_{a_t, 1+n_{a_t}(t-1)} \quad \text{with} \quad n_i(t-1) = \sum_{s=1}^{t-1} \mathbb{I}\{a_s = i\}.$$

In the special case that g_i is monotone nonincreasing for each i there exists a largest interleaving $I^*(g) = I(g, a^*)$, where $a_t^* = \operatorname{argmax}_i g_{a_t, n_{t-1, i}}$. The following lemma follows from the Hardy–Littlewood inequality and we leave the proof as an exercise.

LEMMA 34.3 *If $g_{i1} \leq g_{i2} \leq \dots$ for each i and $\alpha \in (0, 1)$, then*

$$\sum_{t=1}^{\infty} \alpha^t I_t^*(g) = \sup_a \sum_{t=1}^{\infty} \alpha^t I_t(g, a).$$

Proof of Theorem 34.4 Let $\underline{G}(t) = \min_{s \leq t} G^*(S_{A_t}(s))$ and define an increasing sequence of stopping times $(\tau_k)_{k=0}^\infty$ by

$$\tau_0 = 1 \quad \text{and} \quad \tau_{k+1} = \min \{t > \tau_k : A_t \neq A_{\tau_k} \text{ or } \underline{G}(t) < \underline{G}(\tau_k - 1)\}.$$

For the Gittins index policy the τ_{k+1} is exactly the stopping time given in Lemma 34.2. Let $k \in \mathbb{N}$ and abbreviate $i = A_{\tau_k}$. Then

$$\begin{aligned} \mathbb{E} \left[\sum_{t=\tau_k}^{\tau_{k+1}-1} \alpha^t r(S_i(t)) \mid \mathcal{F}_{\tau_k} \right] &= \underline{G}(\tau_k) \mathbb{E} \left[\sum_{t=\tau_k}^{\tau_{k+1}-1} \alpha^t \mid \mathcal{F}_{\tau_k} \right] \text{ a.s.} \\ &= \mathbb{E} \left[\sum_{t=\tau_k}^{\tau_{k+1}-1} \underline{G}(t) \alpha^t \mid \mathcal{F}_{\tau_k} \right] \text{ a.s.,} \end{aligned}$$

where the first equality follows from the definition of the stopping time and Eq. (34.12) and the second because the definition of the stopping time ensures that $\underline{G}(t) = \underline{G}(\tau_k)$ on $\{\tau_k, \dots, \tau_{k+1} - 1\}$. Let S_{iu} be the state of the i th Markov chain when $T_i(t-1) = u$ and $H_{iu} = \min_{v \leq u} G^*(S_{iv})$. The key point is that the distribution of H does not depend on the choice of policy and clearly H_{iu} is monotone nonincreasing in u for each i . Substituting the previous display into the definition of the value function shows that

$$V^{\pi^*} = \mathbb{E} \left[\sum_{k=0}^{\infty} \sum_{t=\tau_k}^{\tau_{k+1}-1} \alpha^t \underline{G}(t) \right] = \mathbb{E} \left[\sum_{t=1}^{\infty} \alpha^t I_t(H, A) \right] = \mathbb{E} \left[\sum_{t=1}^{\infty} \alpha^t I_t^*(H) \right]$$

For policies other than the Gittins policy we note that

$$\mathbb{E} \left[\sum_{t=\tau_k}^{\tau_{k+1}-1} \alpha^t r(S_i(t)) \mid \mathcal{F}_{\tau_k} \right] \leq \mathbb{E} \left[\sum_{t=\tau_k}^{\tau_{k+1}-1} \alpha^t \underline{G}(t) \mid \mathcal{F}_{\tau_k} \right] \text{ a.s.}$$

Summing over all k and taking expectation combined with Lemma 34.3 yields

$$V^\pi \leq \mathbb{E} \left[\sum_{t=1}^{\infty} \alpha^t \underline{G}(t) \right] = \mathbb{E} \left[\sum_{t=1}^n I_t(H, A) \right] \leq \mathbb{E} \left[\sum_{t=1}^n I_t^*(H) \right]. \quad \square$$

34.7 Computing the Gittins index

We describe a simple approach that depends on the state space being finite. References to more general methods are given in the bibliographic remarks. Assume without loss of generality that $\mathcal{S} = \{1, 2, \dots, |\mathcal{S}|\}$ and $\mathcal{G} = 2^{\mathcal{S}}$. The matrix form of the transition kernel is $P \in [0, 1]^{|\mathcal{S}| \times |\mathcal{S}|}$ and is defined by $P_{ij} = \mu(i, \{j\})$. We also let $r \in [0, 1]^{|\mathcal{S}|}$ be the vector of rewards so that $r_i = r(i)$. The standard basis vector is $e_i \in \mathbb{R}^{|\mathcal{S}|}$ and $\mathbf{1} \in \mathbb{R}^{|\mathcal{S}|}$ is the vector with 1 in every coordinate. For $C \subset \mathcal{S}$ we let Q_C be the transition matrix with $(Q_C)_{ij} = P_{ij}\mathbb{I}_C(j)$. For each

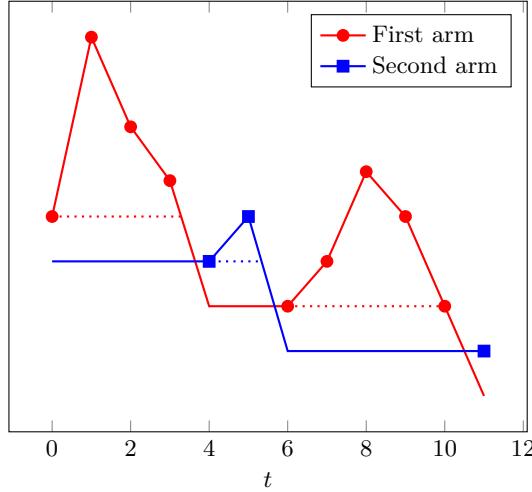


Figure 34.2 The evolution of the fair charge $G^*(S_i(t))$ and prevailing charge $\underline{G}(t)$ for a two-armed bandit. The solid lines indicate the fair charge for each arm, while dotted lines indicate the prevailing charge. Marks indicate which arm is played in each round.

$i \in \mathcal{S}$ our goal is to find

$$G^*(i) = \sup_{\tau > 1} \frac{\mathbb{E}_i \left[\sum_{t=1}^{\tau-1} \alpha^t r(S_t) \right]}{\mathbb{E}_i \left[\sum_{t=1}^{\tau-1} \alpha^t \right]},$$

where \mathbb{E}_i is the expectation with respect to the measure \mathbb{P}_i for which the initial state is $S_1 = i$. The second part of Lemma 34.2 shows that the stopping time $\tau = \min\{t > 1 : G^*(S_t) \leq G^*(i)\}$ attains the supremum in the above display. The set $C_i = \{j : G^*(j) > G^*(i)\}$ is called the continuation region and $S_i = \mathcal{S} \setminus C_i$ is the stopping region. Then the Gittins index can be calculated as

$$G^*(i) = \frac{\mathbb{E}_i \left[\sum_{t=1}^{\tau-1} \alpha^t r(S_t) \right]}{\mathbb{E}_i \left[\sum_{t=1}^{\tau-1} \alpha^t \right]} = \frac{\sum_{t=1}^{\infty} \alpha^t e_i^\top Q_{C_i}^{t-1} r}{\sum_{t=1}^{\infty} \alpha^t e_i^\top Q_{C_i}^{t-1} \mathbf{1}} = \frac{e_i^\top (I - \alpha Q_{C_i})^{-1} r}{e_i^\top (I - \alpha Q_{C_i})^{-1} \mathbf{1}}.$$

All this suggests an induction approach where the Gittins index is calculated for each state in decreasing order of their indices. To get started note that the maximum possible Gittins index is $\max_i r_i$ and that this is achievable for state $i = \operatorname{argmax}_j r_j$ with deterministic stopping time $\tau = 2$. Assume that $G^*(i)$ is known for the k states $C = \{i_1, i_2, \dots, i_k\}$ with the largest indices. Then i_{k+1} is given by

$$i_{k+1} = \operatorname{argmax}_{i \notin C} \frac{e_i^\top (I - \alpha Q_C)^{-1} r}{e_i^\top (I - \alpha Q_C)^{-1} \mathbf{1}}.$$

If Gauss-Jordan elimination is used for matrix inversion, then the computational complexity of this algorithm is $O(|\mathcal{S}|^4)$. A more sophisticated inversion algorithm

would reduce the complexity to $O(|\mathcal{S}|^{3+\varepsilon})$ for some $\varepsilon \leq 0.373$, but these are seldom practical. When α is relatively small the inversion can be replaced by directly calculating the sums to some truncated horizon with little loss in accuracy. There are many other ways to compute the Gittins index with better complexity guarantees. We refer the reader to the bibliographic remarks for references.

34.8 Notes

- 1 An advantage of Bayesian methods is that they automatically and optimally exploit the assumptions. For example, the Bayesian optimal policy for one-armed Bernoulli bandits that we analyzed empirically is essentially the same as its frequentist cousin, but with the tightest possible confidence bounds. This blessing can also be a curse. A policy that exploits its assumptions too heavily can be brittle when those assumptions turn out to be wrong. This can have a devastating effect in bandits where the cost of overly aggressive confidence intervals is large.
- 2 The issue of conditioning on measure zero sets has been described in many places. We do not know of a practical situation where things go really awry. Sensible choices yield sensible posteriors. The curious reader could probably burn a few weeks reading through the literature on the **Borel–Kolmogorov paradox**.
- 3 Economists have long recognized the role of time in the utility people place on rewards. Most people view a promise of pizza a year from today as less valuable than the same pizza tomorrow. Discounting rewards is one way to model this kind of preference. The formal model is credited to renowned American economist Paul Samuelson [1937], who according to Frederick et al. [2002] had serious reservations about both the normative and descriptive value of the model. While discounting is not very common in the frequentist bandit literature, it appears often in reinforcement learning where it offers certain technical advantages [Sutton and Barto, 1998].
- 4 Theorem 34.4 only holds for geometric discounting. If α^t is replaced by $\alpha(t)$ where $\alpha(\cdot)$ is not an exponential, then one can construct Markov chains for which the optimal policy is not an index policy. The intuition behind this result is that when $\alpha(t)$ is not an exponential function, then the Gittins index of an arm can change even in rounds you play a different arm and this breaks the interleaving argument [Berry and Fristedt, 1985].
- 5 We mentioned that computing the Bayesian optimal policy in finite horizon bandits is computationally intractable. But this is not quite true if n is small. For example, when $n = 50$ and $K = 5$ the dynamic program for computing the exact Bayesian optimal policy for Bernoulli noise and Beta prior has approximately 10^{11} states. A big number to be sure, but not so large that the table cannot be stored on disk. And this is without any serious effort to exploit

symmetries. Perhaps for mission-critical applications with small horizon the benefits of exact optimality make the computation worth the hassle.

- 6 The algorithm in Section 34.7 for computing Gittins index is called **Varaiya's algorithm**. In the bibliographic remarks we give some pointers on where to look for more sophisticated methods. The assumption that $|\mathcal{S}|$ is finite is less severe than it may appear. When the discount rate is not too close to 1, then for many problems the Gittins index can be approximated by removing states that are not reachable from the start state before the discounting means they becomes close to irrelevant. When the state space is infinite there is often a topological structure that makes a discretization possible.

34.9 Bibliographical remarks

There are many texts on Bayesian statistics. It's hard not to recommend the book by [Gelman et al. \[2014\]](#) who is one of the main proponents of Bayesian methods. A more philosophical book that takes a foundational look at probability theory from a Bayesian perspective is by [Jaynes \[2003\]](#). The careful definition of the posterior can be found in several places, but the recent book by [Ghosal and van der Vaart \[2017\]](#) does an impeccable job. A worthy mention goes to the article by [Chang and Pollard \[1997\]](#), which uses disintegration to formalise the “private calculations” that probabalists so frequently make before writing everything carefully using Nikodym derivatives and regular versions. Theorem 34.2 is well known. For a simple proof see Theorem 5.3 in the book by [Kallenberg \[2002\]](#). The classic text on optimal stopping is by [Robbins et al. \[1971\]](#), while a more modern text is by [Peskir and Shiryaev \[2006\]](#), which includes a proof of Theorem 34.3 (see Thm. 1.2). For a detailed presentation of exponential families see the book by [Lehmann and Casella \[2006\]](#). We are not aware of a reference for Theorem 34.1, but [Lai \[1987\]](#) has shown that for sufficiently regular priors and noise models the asymptotic Bayesian optimal regret is $\text{BR}_n^* \sim c \log(n)^2$ for some constant $c > 0$ that depends on the prior/model. The Bayesian approach dominated research on bandits from 1960–1980, with Gittins' result (Theorem 34.4) the most celebrated [[Gittins, 1979](#)]. [Gittins et al. \[2011\]](#) has written a whole book on Bayesian bandits. Another book that focusses mostly on the Bayesian problem is by [Berry and Fristedt \[1985\]](#). Although it is now more than thirty years old this book is still a worthwhile read and presents many curious and unintuitive results about exact Bayesian policies. As far as we know the earliest fully Bayesian analysis is by [Bradt et al. \[1956\]](#), who studied the finite horizon Bayesian one-armed bandit problem, essentially writing down the optimal policy using backwards induction as presented here in Section 34.5. For more general approximation results there is the article by [Burnetas and Katehakis \[2003\]](#), which shows that under weak assumptions the Bayesian optimal strategy for one-armed bandits is asymptotically approximated by a retirement policy reminiscent of Eq. (34.10). The very specific approach to approximating the Bayesian strategy for Gaussian

one-armed bandits is by one of the authors [Lattimore, 2016a], where a precise approximation for this special case is also given. There are at least four proofs of Gittins' theorem [Gittins, 1979, Whittle, 1980, Weber, 1992, Tsitsiklis, 1994]. All are summarized in the review by Frostig and Weiss [1999]. There is a line of work on computing and/or approximating the Gittins index, which we cannot do justice to. The approach presented here for finite state spaces is due to Varaiya et al. [1985], but more sophisticated algorithms exist with better guarantees. A nice survey is by Chakravorty and Mahajan [2014], but see also the articles by Chen and Katehakis [1986], Kallenberg [1986], Sonin [2008], Niño-Mora [2011], Chakravorty and Mahajan [2013]. There is also a line of work on approximations of the Gittins index, most of which are based on approximating the discrete time stopping problem with continuous time and applying free boundary methods [Yao, 2006, and references therein]. We mentioned restless bandits in Chapter 31 on nonstationary bandits, but they are usually studied in the Bayesian context Whittle [1988], Weber and Weiss [1990]. The difference is that now the Markov chain for all actions evolve regardless of the action chosen, but the learner only gets to observe the new state for the action they chose.

34.10 Exercises

34.1 Construct an example demonstrating that for some priors over finite-armed stochastic bandits the Bayesian regret is strictly positive: $\inf_{\pi} \text{BR}_n(\pi, \mathbb{Q}) > 0$.

34.2 Evaluate the posteriors for each pair of conjugate priors in Section 34.3.

34.3 Prove Theorem 34.1.

34.4 Prove that the definitions of the Gittins index given in Eq. (34.11) and Eq. (34.12) are equivalent.

34.5 Prove Lemma 34.2.



A proof of this result is given by Frostig and Weiss [1999]. A solution is also available.

34.6 Prove Lemma 34.3.

34.7 This question is about one armed bandits with 1-subgaussian rewards.

- 1 Prove the bound in Eq. (34.6) holds for the retirement policy determined by the stopping time in Eq. (34.7).
- 2 Explain why the policy has bounded regret when $\Delta \geq 0$.

34.8 Reproduce the experimental results in Section 34.5.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

35 Thompson Sampling

“As all things come to an end, even this story, a day came at last when they were in sight of the country where Bilbo had been born and bred, where the shapes of the land and of the trees were as well known to him as his hands and toes.” – [Tolkien \[1937\]](#).

Like Bilbo, as we near the end of the book we return to where it all began, to the first algorithm for bandits proposed by [Thompson \[1933\]](#). The idea is a simple one. Before the game starts the learner chooses a prior over a set of possible bandit environments. In each round the learner samples an environment from the posterior and acts according to the optimal action in that environment. [Thompson](#) only gave empirical evidence (calculated by hand) and focussed on Bernoulli bandits with two arms. Nowadays these limitations have been eliminated and theoretical guarantees have been proven demonstrating the approach is often close to optimal in a wide range of settings. Perhaps more importantly, the resulting algorithms are often quite practical both in terms of computation and empirical performance. The idea of sampling from the posterior and playing the optimal action is called **Thompson sampling** or **posterior sampling**.

The exploration in Thompson sampling comes from the randomization. If the posterior is poorly concentrated, then the fluctuations in the samples are expected to be large and the policy will likely explore. On the other hand, as more data is collected the posterior concentrates towards the true environment and the rate of exploration decreases. We focus our attention on finite-armed stochastic bandits and linear stochastic bandits, but Thompson sampling has been extended to all kinds of models as explained in the bibliographic remarks.



Randomization is crucial for adversarial bandit algorithms and can be useful in stochastic settings (see Chapters [23](#) and [32](#) for examples). We should be wary, however, that there might be a price to pay by injecting variance into our algorithms. What is gained or lost by the randomization in Thompson sampling is still not clear, but we leave this cautionary note as a suggestion to the reader to think about some of the costs and benefits.

35.1 Finite-armed bandits

Recalling very briefly the notation from Section 34.4, let $K > 1$ be the number of arms and $(\mathcal{E}, \mathcal{G}, \mathbb{Q})$ be a probability space where \mathcal{E} is a set of K -armed stochastic bandits and \mathbb{Q} is the prior. For $\nu \in \mathcal{E}$ the distribution on the reward vector in each round is \mathbb{P}_ν on $(\mathbb{R}^K, \mathcal{B}(\mathbb{R}^K))$. As usual we assume that \mathbb{P}_ν is a probability kernel from $(\mathcal{E}, \mathcal{G})$ to $(\mathbb{R}^K, \mathcal{B}(\mathbb{R}^K))$. The reward vector in round t is $X_t \in \mathbb{R}^K$ and the learner observes X_{tA_t} . The posterior after t observations is a random measure \mathbb{Q}_t on $(\mathcal{E}, \mathcal{G})$. The mean of the i th arm in bandit $\nu \in \mathcal{E}$ is $\mu_i(\nu)$.

```

1: Input  $K, \mathcal{E}$  and prior  $\mathbb{Q}$ 
2: for  $t = 1, 2, \dots, n$  do
3:   Compute posterior  $\mathbb{Q}_{t-1}$  based on observed data
4:   Sample  $\nu_t \sim \mathbb{Q}_{t-1}$ 
5:   Choose  $A_t = \operatorname{argmax}_{i \in [K]} \mu_i(\nu_t)$ 
6: end for
```

Bayesian analysis

Thompson sampling has been analyzed in both the frequentist and the Bayesian settings. We start with the latter where the result requires almost no assumptions on the prior. In fact, after one small observation about Thompson sampling, the analysis is almost the same as that of UCB.

THEOREM 35.1 *Let \mathcal{E} a set of 1-subgaussian bandits with K arms and mean rewards bounded in $[0, 1]$ and \mathbb{Q} be a measure on $(\mathcal{E}, \mathcal{G})$ for some σ -algebra \mathcal{G} and π be the policy of Thompson sampling with this prior. Then*

$$\text{BR}_n(\pi, \mathbb{Q}) \leq C \sqrt{Kn \log(n)},$$

where $C > 0$ is a universal constant.

Proof Let \mathbb{P} be the joint measure defined after Eq. (34.2) and ν, A_t and X_t be the coordinate projections given in Eq. (34.3). Expectations are taken with respect to \mathbb{P} . Abbreviate $\hat{\mu}_i = \hat{\mu}_i(\nu)$ and let $A^* = \operatorname{argmax}_{i \in [K]} \mu_i$ be the optimal arm, which depends on ν and is a random variable. For each $t \in [n]$ and $i \in [K]$ let

$$U_t(i) = \text{clip}_{[0,1]} \left(\hat{\mu}_i(t-1) + \sqrt{\frac{2 \log(1/\delta)}{1 \vee T_i(t-1)}} \right),$$

where $\hat{\mu}_i(t-1)$ is the empirical estimate of the reward of arm i after $t-1$ rounds and we assume $\hat{\mu}_i(t-1) = 0$ if $T_i(t-1) = 0$. Let E be the event that for all $t \in [n]$ and $i \in [K]$,

$$|\hat{\mu}_i(t-1) - \mu_i| < \sqrt{\frac{2 \log(1/\delta)}{1 \vee T_i(t-1)}}.$$

In Exercise 35.1 we ask you to prove that $\mathbb{P}(E^c) \leq nK\delta$. Note that $U_t(i)$ is \mathcal{F}_{t-1} -measurable. The Bayesian regret is

$$\text{BR}_n = \mathbb{E} \left[\sum_{t=1}^n (\mu_{A^*} - \mu_{A_t}) \right] = \mathbb{E} \left[\sum_{t=1}^n \mathbb{E} [\mu_{A^*} - \mu_{A_t} \mid \mathcal{F}_{t-1}] \right].$$

The key insight is to notice that the definition of Thompson sampling implies the conditional distributions of A^* and A_t given \mathcal{F}_{t-1} are the same:

$$\mathbb{P}(A^* = \cdot \mid \mathcal{F}_{t-1}) = \mathbb{P}(A_t = \cdot \mid \mathcal{F}_{t-1}) \quad \text{a.s..} \quad (35.1)$$

Using the previous display,

$$\begin{aligned} \mathbb{E} [\mu_{A^*} - \mu_{A_t} \mid \mathcal{F}_{t-1}] &= \mathbb{E} [\mu_{A^*} - U_t(A_t) + U_t(A_t) - \mu_{A_t} \mid \mathcal{F}_{t-1}] \\ &= \mathbb{E} [\mu_{A^*} - U_t(A^*) + U_t(A_t) - \mu_{A_t} \mid \mathcal{F}_{t-1}] \quad (\text{Eq. (35.1)}) \\ &= \mathbb{E} [\mu_{A^*} - U_t(A^*) \mid \mathcal{F}_{t-1}] + \mathbb{E} [U_t(A_t) - \mu_{A_t} \mid \mathcal{F}_{t-1}]. \end{aligned}$$

Using the tower rule for expectation shows that

$$\text{BR}_n = \mathbb{E} \left[\sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \right]. \quad (35.2)$$

On the event E^c the terms inside the expectation are bounded by $2n$ while on the event E the first sum is negative and the second is bounded by

$$\begin{aligned} \mathbb{E} \{E\} \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) &= \mathbb{E} \{E\} \sum_{t=1}^n \sum_{i=1}^K \mathbb{I}\{A_t = i\} (U_t(i) - \mu_i) \\ &\leq \sum_{i=1}^K \sum_{t=1}^n \mathbb{I}\{A_t = i\} \sqrt{\frac{8 \log(1/\delta)}{1 \vee T_i(t-1)}} \leq \sum_{i=1}^K \int_0^{T_i(n)} \sqrt{\frac{8 \log(1/\delta)}{s}} ds \\ &= \sum_{i=1}^K \sqrt{32T_i(n) \log(1/\delta)} \leq \sqrt{32nK \log(1/\delta)}. \end{aligned}$$

The proof is completed by choosing $\delta = n^{-2}$ and the fact that $\mathbb{P}(E^c) \leq nK\delta$. \square

Frequentist analysis

Bounding the frequentist regret of Thompson sampling is significantly more technical than the Bayesian regret. The trouble is the frequentist regret does not have an expectation with respect to the prior, which means that A_t is not conditionally distributed in the same way as the optimal action (which is not random). For brevity we restrict ourselves to the Gaussian case, but other noise models have also been studied as we discuss at the end of the chapter. To make things simple we assume that $A_t = t$ for $t \in [K]$ and subsequently

$$A_t = \operatorname{argmax}_{i \in [K]} \theta_i(t), \quad (35.3)$$

where $\theta_i(t) \sim \mathcal{N}(\hat{\mu}_i(t-1), 1/T_i(t-1))$. Except for the minor detail that we force the algorithm to choose each arm once in the beginning, this policy is derived by

taking an independent Gaussian prior for the mean of each arm and sending the prior variance to infinity.

THEOREM 35.2 *If the algorithm described in Eq. (35.3) is run on Gaussian bandit $\nu \in \mathcal{E}_{\mathcal{N}}^K(1)$, then*

$$R_n \leq C \sum_{i:\Delta_i > 0} \left(\Delta_i + \frac{\log(n)}{\Delta_i} \right),$$

where $C > 0$ is a universal constant. Furthermore, $\limsup_{n \rightarrow \infty} \frac{R_n}{\log(n)} \leq \sum_{i:\Delta_i > 0} \frac{2}{\Delta_i}$.

Proof of Theorem 35.2 Recall the notation from Part II that $\hat{\mu}_{1s}$ is the empirical reward of the first arm after s plays of this arm. As usual we assume without loss of generality that $\mu_1 = \max_i \mu_i$ so that the first arm is optimal. Define $Q_s(\varepsilon) = \mathbb{P}(\theta_1(t) \geq \mu_1 - \varepsilon \mid T_1(t-1) = s)$, which is

$$Q_s(\varepsilon) = \mathbb{P}_{\eta \sim \mathcal{N}(0, 1/s)} (\hat{\mu}_{1s} + \eta \geq \mu_1 - \varepsilon).$$

Let $\varepsilon_1, \dots, \varepsilon_K$ be a sequence of nonnegative constants to be chosen later and define the event $E_i(t) = \{\theta_i(t) \leq \mu_1 - \varepsilon_i\}$. The plan is to bound $\mathbb{E}[T_i(n)]$ for each suboptimal arm i and then apply Lemma 4.2. We start with a straightforward decomposition.

$$\begin{aligned} \mathbb{E}[T_i(n)] &= \mathbb{E} \left[\sum_{t=1}^n \mathbb{I}\{A_t = i\} \right] \\ &= \mathbb{E} \left[\sum_{t=1}^n \mathbb{I}\{A_t = i, E_i(t)\} \right] + \mathbb{E} \left[\sum_{t=1}^n \mathbb{I}\{A_t = i, E_i^c(t)\} \right]. \end{aligned} \quad (35.4)$$

The second sum on the right-hand side is the easy term. Essentially, if $T_i(t-1)$ is large enough, then the probability of $E_i^c(t)$ is unlikely to be very large. We leave it to the reader in Exercise 35.3 to prove for some universal constant $C > 0$ that

$$\mathbb{E} \left[\sum_{t=1}^n \mathbb{I}\{A_t = i, E_i^c(t)\} \right] \leq C \left(1 + \frac{1}{(\Delta_i - \varepsilon_i)^2} \right). \quad (35.5)$$

The next step is the novel part of the analysis, which bounds the conditional probability that suboptimal arm i is played in round t in terms of the probability of playing the optimal arm. Let $A'_t = \operatorname{argmax}_{i \neq 1} \theta_i(t)$. Then for any $i > 1$,

$$\begin{aligned} \mathbb{P}(A_t = i, E_i(t) \mid \mathcal{F}_{t-1}) &\geq \mathbb{P}(A'_t = i, E_i(t), \theta_1(t) \geq \mu_1 - \varepsilon_i \mid \mathcal{F}_{t-1}) \\ &= \mathbb{P}(\theta_1(t) \geq \mu_1 - \varepsilon_i \mid \mathcal{F}_{t-1}) \mathbb{P}(A'_t = i, E_i(t) \mid \mathcal{F}_{t-1}) \\ &\geq \frac{Q_{T_1(t-1)}(\varepsilon_i)}{1 - Q_{T_1(t-1)}(\varepsilon_i)} \mathbb{P}(A_t = i, E_i(t) \mid \mathcal{F}_{t-1}), \end{aligned} \quad (35.6)$$

where in the first equality we used the fact that $\theta_1(t)$ is conditionally independent

of A'_t and $E_i(t)$ given \mathcal{F}_{t-1} . In the second inequality we used the definition of $Q_{T_1(t-1)}(\varepsilon_i) = \mathbb{P}(\theta_1(t) \geq \mu_1 - \varepsilon_i \mid \mathcal{F}_{t-1})$ and the fact that

$$\mathbb{P}(A_t = i, E_i(t) \mid \mathcal{F}_{t-1}) \leq (1 - \mathbb{P}(\theta_1(t) \geq \mu_1 - \varepsilon_i \mid \mathcal{F}_{t-1}))\mathbb{P}(A'_t = i, E_i(t) \mid \mathcal{F}_{t-1}),$$

which is true because $\{A_t = i, E_i(t)\} \subseteq \{A'_t = i, E_i(t)\} \cap \{\theta_1(t) \leq \mu_1 - \varepsilon_i\}$ and the two intersected events are conditionally independent given \mathcal{F}_{t-1} . Therefore using Eq. (35.6) we have

$$\begin{aligned}\mathbb{P}(A_t = i, E_i(t) \mid \mathcal{F}_{t-1}) &\leq \left(\frac{1}{Q_{T_1(t-1)}(\varepsilon_i)} - 1 \right) \mathbb{P}(A_t = 1, E_i(t) \mid \mathcal{F}_{t-1}) \\ &\leq \left(\frac{1}{Q_{T_1(t-1)}(\varepsilon_i)} - 1 \right) \mathbb{P}(A_t = 1 \mid \mathcal{F}_{t-1}).\end{aligned}$$

Substituting this into the first term in Eq. (35.4) leads to

$$\begin{aligned}\mathbb{E} \left[\sum_{t=1}^n \mathbb{I}\{A_t = i, E_i(t)\} \right] &\leq \mathbb{E} \left[n \wedge \sum_{t=1}^n \left(\frac{1}{Q_{T_1(t-1)}(\varepsilon_i)} - 1 \right) \mathbb{P}(A_t = 1 \mid \mathcal{F}_{t-1}) \right] \\ &= \mathbb{E} \left[n \wedge \sum_{t=1}^n \left(\frac{1}{Q_{T_1(t-1)}(\varepsilon_i)} - 1 \right) \mathbb{I}\{A_t = 1\} \right] \\ &\leq \mathbb{E} \left[n \wedge \sum_{s=1}^n \left(\frac{1}{Q_s(\varepsilon_i)} - 1 \right) \right] \\ &\leq \mathbb{E} \left[\sum_{s=1}^n \left(n \wedge \left(\frac{1}{Q_s(\varepsilon_i)} - 1 \right) \right) \right].\end{aligned}\tag{35.7}$$

where in the second last step we used the fact that $T_1(t-1) = s$ is only possible for one round where $A_t = 1$. At last we have decoupled all the dependencies between the arms and reduced the problem to studying the right-hand side of Eq. (35.7). We will shortly show that for any $\gamma \in (0, 1)$,

$$\sum_{s=1}^n \mathbb{E} \left[n \wedge \left(\frac{1}{Q_s(\varepsilon_i)} - 1 \right) \right] \leq \frac{8 \log \left(e + \frac{8}{\varepsilon_i \gamma^2} \right)}{\varepsilon_i^2 \gamma^2} + \frac{2 \log(n+1)}{\varepsilon_i^2 (1-\gamma)}. \tag{35.8}$$

The theorem follows from the claim and the standard regret decomposition (Lemma 4.2) and by choosing $\varepsilon_i = (1-\gamma)\Delta_i$, where the finite-time result follows with $\gamma = 1/2$ and the asymptotic result with $\gamma = \log^{-1/8}(n)$. The proof of the claim in Eq. (35.8) is a bit of a slog. Let

$$F_s(x) = \sqrt{\frac{s}{2\pi}} \int_{-\infty}^x \exp(-sy^2/2) dy$$

be the cumulative distribution function for a Gaussian with zero mean and

variance $1/s$. Then $Q_s(\varepsilon) = 1 - F_s(\mu_1 - \hat{\mu}_{1s} - \varepsilon)$. Therefore

$$\begin{aligned}\mathbb{E} \left[n \wedge \left(\frac{1}{Q_s(\varepsilon)} - 1 \right) \right] &= \int_0^n \mathbb{P} \left(\frac{1}{Q_s(\varepsilon)} - 1 \geq x \right) dx \\ &= \int_0^n \mathbb{P} \left(F_s(\mu_1 - \hat{\mu}_{1s} - \varepsilon) \geq \frac{x}{1+x} \right) dx \\ &= \int_0^n \mathbb{P} \left(\mu_1 - \hat{\mu}_{1s} - \varepsilon \geq F_s^{-1}(x/(1+x)) \right) dx \\ &= \int_0^n (1 - F_s(\varepsilon + F_s^{-1}(x/(1+x)))) dx,\end{aligned}$$

where in the last line we used the fact that $\mu_1 - \hat{\mu}_{1s}$ is Gaussian with zero mean and variance $1/s$. By Theorem 5.1 it holds that $F_s(-\varepsilon\gamma/2) \leq \exp(-s\varepsilon^2\gamma^2/8)$. Therefore if $x/(1+x) \geq u = \exp(-s\varepsilon^2\gamma^2/8)$, then $F_s^{-1}(x/(1+x)) \geq -\varepsilon\gamma/2$. Abbreviating $g_s(x) = 1 - F_s(\varepsilon + F_s^{-1}(x/(1+x)))$ we see that for $x \geq u/(1-u)$,

$$\begin{aligned}g_s(x) &= \int_x^\infty g'_s(y) dy = \int_x^\infty \frac{\exp\left(-\frac{s\varepsilon^2+2s\varepsilon F_s^{-1}(y/(1+y))}{2}\right)}{(1+y)^2} dy \\ &\leq \int_x^\infty \frac{\exp\left(-\frac{s\varepsilon^2+2s\varepsilon F_s^{-1}(x/(1+x))}{2}\right)}{(1+y)^2} dy \leq \int_x^\infty \frac{\exp\left(-\frac{s(1-\gamma)\varepsilon^2}{2}\right)}{(1+y)^2} dy \\ &= \frac{\exp\left(-\frac{s(1-\gamma)\varepsilon^2}{2}\right)}{1+x}.\end{aligned}$$

From its definition it is easily seen that $g_s(x) \leq 1 - x/(1+x) = 1/(1+x)$ so that by splitting the integral we have

$$\begin{aligned}\mathbb{E} \left[n \wedge \left(\frac{1}{Q_s(\varepsilon)} - 1 \right) \right] &= \int_0^n g_s(x) dx \\ &\leq \int_0^{u/(1-u)} \frac{dx}{1+x} + \exp\left(-\frac{s(1-\gamma)\varepsilon^2}{2}\right) \int_{x_0}^n \frac{dx}{1+x} \\ &\leq \log\left(\frac{1}{1 - \exp\left(-\frac{s\gamma^2\varepsilon^2}{8}\right)}\right) + \exp\left(-\frac{s(1-\gamma)\varepsilon^2}{2}\right) \log(n+1). \quad (35.9)\end{aligned}$$

We make use of the following facts:

$$\sum_{s=1}^{\infty} \exp(-sp) \leq \frac{1}{p} \quad \text{and} \quad \sum_{s=1}^{\infty} \log\left(\frac{1}{1 - \exp(-sp)}\right) \leq \frac{\log(e+1/p)}{p}.$$

Summing Eq. (35.9) over s and applying the facts yields the proof of Eq. (35.8):

$$\sum_{s=1}^n \mathbb{E} \left[n \wedge \left(\frac{1}{Q_s(\varepsilon_i)} - 1 \right) \right] \leq \frac{8 \log\left(e + \frac{8}{\varepsilon_i^2\gamma^2}\right)}{\varepsilon_i^2\gamma^2} + \frac{2 \log(n+1)}{\varepsilon_i^2(1-\gamma)}. \quad \square$$

35.2 Linear bandits

While the advantages of Thompson sampling in finite-armed bandits are relatively limited, in the linear setting there is much to be gained, both in terms of computation and empirical performance. Let \mathcal{E} be the set of Gaussian linear bandits with a fixed action-set $\mathcal{A} \subset \mathbb{R}^d$. A Gaussian linear bandit is characterized by its mean vector $\theta \in \mathbb{R}^d$ and the reward after taking action A_t in round t is

$$X_t = \langle A_t, \theta \rangle + \eta_t,$$

where η_1, \dots, η_n is a sequence of independent standard Gaussian random variables. A prior corresponds to choosing a measure on \mathbb{R}^d . An advantage of Thompson sampling relative to optimistic linear bandit algorithms is that the optimization problem for selecting the action no longer requires optimizing over a confidence ellipsoid. There are many cases where this makes a significant difference. For example, if \mathcal{A} is convex, then Thompson sampling can often be computed efficiently, which is not generally the case for the optimistic linear bandit algorithms in Chapter 19.

```

1: Input Prior  $\mathbb{Q}$  and action-set  $\mathcal{A}$ 
2: for  $t \in 1, \dots, n$  do
3:   Sample  $\theta_t$  from the posterior
4:   Choose  $A_t = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \theta_t \rangle$ 
5: end for
```

Algorithm 21: Thompson sampling for linear bandits

The Bayesian regret is controlled using the techniques from the previous section in combination with the concentration analysis in Chapter 20. A frequentist analysis is also possible under slightly unsatisfying assumptions, which we discuss in the notes and bibliographic remarks.

THEOREM 35.3 *Assume that $\|\theta\|_2 \leq S$ with \mathbb{Q} -probability one and $\sup_{a \in \mathcal{A}} \|a\|_2 \leq L$ and $\sup_{a \in \mathcal{A}} \langle a, \theta \rangle \leq 1$ with \mathbb{Q} -probability one. Then the Bayesian regret of Algorithm 21 is bounded by*

$$\text{BR}_n \leq 2 + 2\sqrt{2dn\beta^2 \log\left(1 + \frac{nS^2L^2}{d}\right)},$$

$$\text{where } \beta = 1 + \sqrt{2\log(n) + d\log\left(1 + \frac{nS^2L^2}{d}\right)}.$$

Proof We apply the same technique as the proof of Theorem 35.1. Define upper confidence bound $U_t : \mathcal{A} \rightarrow \mathbb{R}$ by

$$U_t(a) = \langle \hat{\theta}_{t-1}, a \rangle + \beta \|a\|_{V_{t-1}^{-1}}, \quad \text{where } V_t = \frac{I}{\mathbb{E}[\|\theta\|_2^2]} + \sum_{s=1}^t A_s A_s^\top.$$

By Theorem 20.2, $\mathbb{P}(\text{exists } t \leq n : \|\hat{\theta} - \theta\|_{V_t} \geq \beta) \leq 1/n$. Let E_t be the event that $\|\hat{\theta}_{t-1} - \theta\|_{V_{t-1}} < \beta$ and $E = \bigcap_{t=1}^n E_t$ and $A^* = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \theta \rangle$, which is a random variable in this setting because θ is random. Then

$$\begin{aligned} \text{BR}_n &= \mathbb{E} \left[\sum_{t=1}^n \langle A^* - A_t, \theta \rangle \right] \\ &= \mathbb{E} \left[\mathbb{I}_{E^c} \sum_{t=1}^n \langle A^* - A_t, \theta \rangle \right] + \mathbb{E} \left[\mathbb{I}_E \sum_{t=1}^n \langle A^* - A_t, \theta \rangle \right] \\ &\leq 2 + \mathbb{E} \left[\mathbb{I}_E \sum_{t=1}^n \langle A^* - A_t, \theta \rangle \right] \\ &\leq 2 + \mathbb{E} \left[\sum_{t=1}^n \mathbb{I}_{E_t} \langle A^* - A_t, \theta \rangle \right]. \end{aligned} \quad (35.10)$$

As before, $\mathbb{P}(A^* = \cdot \mid \mathcal{F}_{t-1}) = \mathbb{P}(A_t = \cdot \mid \mathcal{F}_{t-1})$, which means the second term in the above display is bounded by

$$\begin{aligned} \mathbb{E}_{t-1} [\mathbb{I}_{E_t} \langle A^* - A_t, \theta \rangle] &= \mathbb{E}_{E_t} \mathbb{E}_{t-1} [\langle A^*, \theta \rangle - U_t(A^*) + U_t(A_t) - \langle A_t, \theta \rangle] \\ &\leq \mathbb{E}_{E_t} \mathbb{E}_{t-1} [U_t(A^*) - \langle A_t, \theta \rangle] \\ &\leq \mathbb{E}_{E_t} \mathbb{E}_{t-1} [\langle A_t, \hat{\theta}_{t-1} - \theta \rangle] + \beta \|A_t\|_{V_t^{-1}} \\ &\leq \mathbb{E}_{E_t} \mathbb{E}_{t-1} [\|A_t\|_{V_t^{-1}} \|\hat{\theta}_{t-1} - \theta\|_{V_t}] + \beta \|A_t\|_{V_t^{-1}} \\ &\leq 2\beta \|A_t\|_{V_t^{-1}}. \end{aligned}$$

Substituting into the second term of Eq. (35.10),

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^n \mathbb{I}_{E_t} \langle A^* - A_t, \theta \rangle \right] &\leq 2\mathbb{E} \left[\beta \sum_{t=1}^n (1 \wedge \|A_t\|_{V_t^{-1}}) \right] \\ &\leq 2 \sqrt{n \mathbb{E} \left[\beta^2 \sum_{t=1}^n (1 \wedge \|A_t\|_{V_t^{-1}}^2) \right]} \quad (\text{Cauchy-Schwartz}) \\ &\leq 2 \sqrt{2dn \mathbb{E} \left[\beta^2 \log \left(1 + \frac{nS^2L^2}{d} \right) \right]}. \quad (\text{Lemma 19.1}) \end{aligned}$$

Putting together the pieces shows that

$$\text{BR}_n \leq 2 + 2 \sqrt{2dn \beta^2 \log \left(1 + \frac{nS^2L^2}{d} \right)}. \quad \square$$

Computation

An implementation of Thompson sampling for linear bandits needs to (a) sample θ_t from the posterior and (b) find the optimal action for the sampled parameter:

$$A_t = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \theta_t \rangle.$$

For some priors and noise models sampling from the posterior is straightforward. The most notable case is when \mathbb{Q} is a multivariate Gaussian. More generally there is a large literature devoted to numerical methods for sampling from posterior distributions. Have sampled θ_t , the optimization problem of finding A_t is a linear optimization problem. Compare this to LinUCB, which needs to solve

$$A_t = \operatorname{argmax}_{a \in \mathcal{A}} \operatorname{argmax}_{\tilde{\theta} \in \mathcal{C}} \langle a, \tilde{\theta} \rangle,$$

which for large or continuous action-sets is usually much harder computationally.

35.3 Information theoretic analysis

The analysis in the previous sections mirrored those for the frequentist algorithms in Part II. Here we showcase a different approach that relies exclusively on information theory. The argument is based on the observation that for bandits with K arms at most $\log(K)$ nats are needed to code the identity of the optimal arm, which means the total information gain about this quantity is bounded. For many policies one can prove a relationship between the information gain about the optimal arm and the expected regret, which in combination with the previous observation leads to a bound on the regret. This analysis is all the more striking because the assumption that the bandits are (stationary) stochastic can be relaxed as we discuss in the notes.

A few more definitions from information theory are needed. Let X be a discrete random variable on probability space $(\Omega, \mathcal{G}, \mathbb{P})$. Recall from Chapter 14 that the entropy of X is defined by

$$H(X) = \sum_{x \in \text{range}(X)} \mathbb{P}(X = x) \log \left(\frac{1}{\mathbb{P}(X = x)} \right).$$

We also need the **conditional entropy**. Let $\mathcal{F} \subset \mathcal{G}$ be a σ -algebra. Then

$$H(X | \mathcal{F}) = \mathbb{E} \left[\sum_{x \in \text{range}(X)} \mathbb{P}(X = x | \mathcal{F}) \log \left(\frac{1}{\mathbb{P}(X = x | \mathcal{F})} \right) \right].$$

A little confusingly, the conditional entropy is *not* a random variable. Perhaps a better nomenclature would have been the expected conditional entropy. The entropy of random variable X is a measure of the amount of information in X while the conditional entropy given \mathcal{F} is the expected amount of information required to encode X having observed the information in \mathcal{F} . The **mutual information** between X and \mathcal{F} is the difference between the entropy and the conditional entropy:

$$I(X; \mathcal{F}) = H(X) - H(X | \mathcal{F}).$$

The mutual information is always nonnegative, which should not be surprising because the information remaining in X can only decrease as more information

is observed. Another name for the mutual information is **information gain**. We use these forms when the underlying measure is clear from context. If this is not the case, then the measure is shown in the subscript: $H(X) = H_{\mathbb{P}}(X)$. The following lemmas provide a chain rule for the mutual information and a simple connection to the relative entropy. The proofs are definitional and are left as exercises.

LEMMA 35.1 *Let X be a random variable on $(\Omega, \mathcal{G}, \mathbb{P})$ and $(\mathcal{F}_t)_{t=0}^n$ a filtration of \mathcal{G} with $\mathcal{F}_0 = \{\emptyset, \mathcal{F}\}$ and $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot | \mathcal{F}_t)$. Then*

$$\mathbb{E} \left[\sum_{t=1}^n I_{\mathbb{P}_{t-1}}(X; \mathcal{F}_t) \right] = I_{\mathbb{P}}(X; \mathcal{F}_n).$$

LEMMA 35.2 *Let X and Y be random variables on probability space $(\Omega, \mathcal{G}, \mathbb{P})$. If X is discrete and $I(X; Y)$ exists, then*

$$I(X; Y) = \mathbb{E} [D(\mathbb{P}_{Y|X}, \mathbb{P}_Y)],$$

where $\mathbb{P}_{Y|X}$ is the random measure on $(\Omega, \sigma(Y))$ such that $\mathbb{P}_{Y|X}(A) = \mathbb{P}(Y \in A | X)$ almost surely.

We now present an elegant result connecting the Bayesian regret of any policy and the information gain. Recall that $\mathcal{F}_t = \sigma(A_1, X_{1A_1}, \dots, A_t, X_{tA_t})$ and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_t]$ and $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot | \mathcal{F}_t)$ and abbreviate $H_t(\cdot) = H_{\mathbb{P}_t}(\cdot)$ and $I_t(\cdot; \cdot) = I_{\mathbb{P}_t}(\cdot; \cdot)$. Define random variable Γ_t to be the ratio of the squared expected Bayesian instantaneous regret and the information gain about the optimal arm.

$$\Gamma_t = \frac{(\mathbb{E}_{t-1}[X_{tA^*} - X_{tA_t}])^2}{I_{t-1}(A^*; (A_t, X_{tA_t}))}. \quad (35.11)$$

THEOREM 35.4 *Suppose that $\Gamma_t \leq \bar{\Gamma}$ almost surely for all $t \in [n]$. Then*

$$\text{BR}_n \leq \sqrt{n\bar{\Gamma}H(A^*)}.$$

Proof By the definitions of the regret and Γ_t in Eq. (35.11) and Cauchy-Schwartz we have

$$\begin{aligned} \text{BR}_n &= \mathbb{E} \left[\sum_{t=1}^n (X_{tA^*} - X_{tA_t}) \right] = \mathbb{E} \left[\sum_{t=1}^n \mathbb{E}_{t-1}[X_{tA^*} - X_{tA_t}] \right] \\ &= \mathbb{E} \left[\sum_{t=1}^n \sqrt{I_{t-1}(A^*; (A_t, X_{tA_t}))\Gamma_t} \right] \leq \sqrt{\bar{\Gamma}} \mathbb{E} \left[\sum_{t=1}^n \sqrt{I_{t-1}(A^*; (A_t, X_{tA_t}))} \right] \\ &\leq \sqrt{n\bar{\Gamma} \mathbb{E} \left[\sum_{t=1}^n I_{t-1}(A^*; (A_t, X_{tA_t})) \right]} \leq \sqrt{n\bar{\Gamma}H(A^*)}, \end{aligned}$$

where the last inequality follows from Lemma 35.1. \square



Theorem 35.4 holds for any policy and clearly illustrates the ‘learn something’ or ‘suffer no regret’ argument that appeared in the analyses of so many algorithms. If the ratio of regret relative to information is large, then a policy could suffer high regret. By contrast, policies for which the regret-information ratio is small will enjoy strong regret guarantees.

A combination of Theorem 35.4 and an almost sure bound on Γ_t can lead to nearly optimal bounds on the Bayesian regret of Thompson sampling for finite-armed and linear bandits. We present only the finite-armed case.

LEMMA 35.3 *If $X_{ti} \in [0, 1]$ almost surely for all $t \in [n]$ and $i \in [K]$ and A_t is chosen by Thompson sampling using any prior, then $\Gamma_t \leq \frac{K}{2}$ almost surely.*

Before the proof of the lemma we note the consequences. Let \mathcal{E} be a set of finite-armed bandits with K arms and rewards in $[0, 1]$. Then Thompson sampling with any prior has its Bayesian regret bounded by

$$\text{BR}_n \leq \sqrt{\frac{Kn \log(K)}{2}}. \quad (35.12)$$

Proof of Lemma 35.3 To avoid clutter we drop all subscripts on t . By the chain rule for mutual information we have

$$I(A^*; X_A, A) = I(A^*; A) + I(A^*; X_A | A) \quad (35.13)$$

The first term in the above display vanishes because A and A^* are independent under \mathbb{P} (Exercise 35.5).

$$\begin{aligned} I(A^*; X_A | A) &= \sum_{i=1}^K \mathbb{P}(A = i) I(A^*; X_i) \\ &= \sum_{i=1}^K \mathbb{P}(A = i) \sum_{j=1}^K \mathbb{P}(A^* = j) D(\mathbb{P}_{X_i | A^* = j}, \mathbb{P}_{X_i}) \\ &\geq 2 \sum_{i=1}^K \sum_{j=1}^K \mathbb{P}(A = i) \mathbb{P}(A^* = j) (\mathbb{E}[X_i | A^* = j] - \mathbb{E}[X_i])^2 \\ &\geq 2 \sum_{i=1}^K \mathbb{P}(A = i)^2 (\mathbb{E}[X_i | A^* = i] - \mathbb{E}[X_i])^2 \\ &\geq \frac{2}{K} \left(\sum_{i=1}^K \mathbb{P}(A = i) (\mathbb{E}[X_i | A^* = i] - \mathbb{E}[X_i]) \right)^2, \end{aligned}$$

where the first equality follows by Eq. (35.13) and the second by Lemma 35.2. The first inequality follows from Pinsker’s inequality (Eq. 14.8), the result in Exercise 14.1 and the assumption that the rewards lie in $[0, 1]$. The second inequality follows by dropping cross terms and the third by Cauchy-Schwartz. The result follows by rearranging the above display. \square

35.4 Notes

- 1 Thompson sampling is known to be asymptotically optimal in a variety of settings. Most notably when the noise model follows a single-parameter exponential family and the prior is chosen appropriately [Kaufmann et al., 2012b, Korda et al., 2013]. Unfortunately Thompson sampling is not a golden bullet. The linear variant in Section 35.2 is not asymptotically optimal by the same argument we presented for optimism in Chapter 25. Characterizing the conditions under which Thompson sampling is close to optimal remains an open challenge.
- 2 For the Gaussian noise model it is known that Thompson sampling is not minimax optimal. Its worst case regret is $R_n = \Theta(\sqrt{nK \log(K)})$ [Agrawal and Goyal, 2013a].
- 3 An alternative to sampling from the posterior is to choose in each round the arm that maximizes a **Bayesian upper confidence bound**, which is a quantile of the posterior. The resulting algorithm is called **BayesUCB** and has excellent empirical and theoretical guarantees [Kaufmann et al., 2012a, Kaufmann, 2018].
- 4 The prior has a significant effect on the performance of Thompson sampling. In classical Bayesian statistics a poorly chosen prior is quickly washed away by data. This is not true in bandits because if the prior underestimates the quality of an arm, then Thompson sampling may never play that arm with high probability and no data is ever observed. We ask you to explore this situation in Exercise 35.9.
- 5 An instantiation of Thompson sampling for linear bandits is known to enjoy near-optimal frequentist regret. In each round the algorithm samples $\theta_t \sim \mathcal{N}(\hat{\theta}_{t-1}, rV_{t-1})$, where $r = \Theta(d)$ is a constant and

$$V_t = I + \sum_{s=1}^t A_s A_s^\top \quad \text{and} \quad \hat{\theta}_t = V_t^{-1} \sum_{s=1}^t X_s A_s .$$

Then $A_t = \operatorname{argmax}_{a \in \mathcal{A}} \langle \theta_t, a \rangle$. This corresponds to assuming the noise is Gaussian with variance r and choosing prior $\mathbb{Q} = \mathcal{N}(0, I)$. Provided the rewards are conditionally 1-subgaussian, the frequentist regret of this algorithm is $R_n = \tilde{O}(d^{3/2}\sqrt{n})$, which is worse than LinUCB by a factor of \sqrt{d} . The increased regret is caused by the choice of noise model, which assumes the variance is $r = \Theta(d)$ rather than $r = 1$. The reason to do this comes from the analysis, which works by showing the algorithm is ‘optimistic’ with reasonable probability. It is not known whether or not this is necessary or an artifact of the analysis. Empirical evidence suggests that $r = 1$ leads to improved performance.

- 6 A more generic view of Thompson sampling is via the idea of perturbations. The **follow the perturbed leader** algorithm chooses in each round the action

$$A_t = \operatorname{argmax}_{i \in [K]} (\hat{\mu}_i(t-1) + \eta_{it}) ,$$

where $\eta_{1t}, \dots, \eta_{Kt}$ is a sequence of independent random variables. In many cases Thompson sampling is hard to analyze because the variance of the randomization is not quite sufficient to prove the optimal arm is optimistic with sufficiently large probability. By sacrificing the Bayesian viewpoint one can sometimes derive a similar algorithm for which the analysis is more straightforward.

- 7 The analysis in Section 35.3 can be generalized to structured settings such as linear bandits [Russo and Roy, 2016]. For linear bandits with an infinite action set the entropy of the optimal action may be infinite. The analysis can be corrected in this case by discretizing the action-set and comparing to a near-optimal action. This leads to a tradeoff between the fineness of the discretization and its size. The algorithm does not depend on the discretization. The reader is referred to the recent article by Dong and Roy [2018]. The assumption of bounded rewards in Lemma 35.3 can be relaxed to a subgaussian assumption. For details see the paper by Russo and Roy [2016].
- 8 Nowhere in the proofs of Theorem 35.4 and Lemma 35.3 did we use the fact that bandits in \mathcal{E} are stochastic. Let $\mathcal{E} = [0, 1]^{nK}$ and \mathbb{Q} be a prior probability measure on $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$. We view elements of $\nu \in \mathcal{E}$ as oblivious ‘adversarial’ bandits, which are really just sequences of reward vectors. Let $(X_{ti})_{ti}$ be a sequence of reward vectors sampled from \mathbb{Q} . The optimal action in hindsight is

$$A^* = \operatorname{argmax}_{i \in [K]} \sum_{t=1}^n X_{ti} .$$

The posterior in round t is $\mathbb{Q}_t = \mathbb{Q}(\cdot | A_1, X_{1A_1}, \dots, A_t, X_{tA_t})$. Then in round t Thompson sampling samples $\nu \sim \mathbb{Q}_{t-1}$ and chooses $A_t = \operatorname{argmax}_{i \in [K]} \sum_{t=1}^n \nu_{ti}$. Repeating the analysis in Section 35.3 shows that

$$\mathbb{E} \left[\sum_{t=1}^n X_{tA^*} - X_{tA_t} \right] \leq \sqrt{nK \log(K)/2} .$$

The calculation even works for non-oblivious adversaries, provided of course that X_{tA_t} only depends on $A_1, X_1 \dots, A_{t-1}, X_{t-1}$.

$$\mathbb{P} = \mathbb{Q} \otimes \mathbb{P}_\nu$$

- 9 The previous note highlights a connection between Bayesian regret and the minimax regret in adversarial bandits. First notice that Recall an adversarial Bernoulli finite-armed bandit is a matrix (x_{ti}) with $x_{ti} \in \{0, 1\}$ for all $t \in [n]$ and $i \in [K]$. Let \mathcal{E} be the set of all adversarial bandits and Π the set of all randomized policies and \mathcal{Q} be the set of all distributions on \mathcal{E} . Then by the

minimax theorem of [Sion \[1958\]](#),

$$\begin{aligned} R_n^* &= \min_{\pi \in \Pi} \max_{(x_{ti}) \in \mathcal{E}} \mathbb{E}_{\pi, x} \left[\max_{i \in [K]} \sum_{t=1}^n (x_{ti} - x_{tA_t}) \right] \\ &= \max_{Q \in \mathcal{Q}} \min_{\pi \in \Pi} \underbrace{\mathbb{E}_{x \sim Q} \left[\mathbb{E}_{\pi, x} \left[\sum_{t=1}^n (x_{ti} - x_{tA_t}) \right] \right]}_{\text{Bayesian regret}}. \end{aligned}$$

The consequence is that if the regret of the Bayesian optimal algorithm is bounded by B for all priors Q , then the minimax adversarial regret is bounded by B . By Eq. (35.12) we can conclude there exists an adversarial bandit algorithm with worst case regret at most $\sqrt{Kn \log(K)/2}$, which can be strengthened using the result in [Exercise 35.2](#). Of course we already knew these things, but the approach has applications in more sophisticated settings. The most notable example being the first near-optimal analysis for adversarial convex bandits [[Bubeck et al., 2015a](#), [Bubeck and Eldan, 2016](#)]. The main disadvantage is that uniform bounds on the Bayesian regret implies existence of a single algorithm with small minimax adversarial regret, but the result is nonconstructive.

- 10 The information-theoretic ideas in Section 35.3 suggest that rather than sampling A_t from the posterior on A^* , one can sample A_t from the distribution minimizing Eq. (35.11). Specifically, A_t is sampled from distribution π_t on $[K]$ where

$$\pi_t = \operatorname{argmin}_\pi \frac{\left(\sum_{i=1}^K \pi(i) (\mathbb{E}_{t-1}[X_{tA^*} - X_{ti}]) \right)^2}{\sum_{i=1}^K \pi(i) I_{t-1}(A^*; X_{ti} \mid A_t = i)}.$$

The resulting policy is called **Information Directed Sampling**. Bayesian regret analysis for this algorithm follows along similar lines as what was presented in Section 35.3. See the paper by [Russo and Roy \[2014a\]](#) for more details or [Exercise 35.7](#).

35.5 Bibliographic remarks

Thompson sampling has the honor of being the first bandit algorithm and is named after its inventor [[Thompson, 1933](#)], who considered the Bernoulli case with two arms. [Thompson](#) provided no theoretical guarantees, but argued intuitively and gave hand-calculated empirical analysis. It would be wrong to say that Thompson sampling was entirely ignored, but its popularity soared when a large number of authors independently rediscovered the article/algorithm [[Granmo, 2010](#), [Ortega and Braun, 2010](#), [Graepel et al., 2010](#), [Chapelle and Li, 2011](#), [May et al., 2012](#)]. The surge in interest was mostly empirical, but theoreticians followed soon with regret guarantees. For the frequentist analysis we followed the

proofs by [Agrawal and Goyal \[2013a, 2012\]](#), but the setting is slightly different. We presented results for the ‘realizable’ case where the payoff distributions are actually Gaussian, while [Agrawal and Goyal](#) use the same algorithm but prove bounds for rewards bounded in $[0, 1]$. [Agrawal and Goyal \[2013a\]](#) also analyze the Beta/Bernoulli variant of Thompson sampling, which for rewards in $[0, 1]$ is asymptotically optimal in the same way as KL-UCB (see Chapter 10). This result was simultaneously obtained by [Kaufmann et al. \[2012b\]](#), who later showed that for appropriate priors asymptotic optimality holds for single parameter exponential families [\[Korda et al., 2013\]](#). For Gaussian bandits with unknown mean and variance Thompson sampling is asymptotically optimal for some priors, but not others – even quite natural ones [\[Honda and Takemura, 2014\]](#). The Bayesian analysis of Thompson sampling based on confidence intervals is due to [Russo and Roy \[2014b\]](#) while the information-theoretic argument is by [Russo and Roy \[2014a, 2016\]](#). Recently the idea has been applied to a wide range of bandit settings [\[Kawale et al., 2015, Agrawal et al., 2017\]](#) and reinforcement learning [\[Osband et al., 2013, Gopalan and Mannor, 2015, Leike et al., 2016, Kim, 2017\]](#). The BayesUCB algorithm is due to [Kaufmann et al. \[2012a\]](#) with improved analysis and results by [Kaufmann \[2018\]](#). The frequentist analysis of Thompson sampling for linear bandits is by [Agrawal and Goyal \[2013b\]](#) with refined analysis by [Abeille and Lazaric \[2017a\]](#) and a spectral version by [Kocák et al. \[2014\]](#). There is a tutorial on Thompson sampling by [Russo et al. \[2017\]](#) that focuses mostly on applications and computational issues.

35.6 Exercises

35.1 Consider the event E defined in Theorem 35.1 and prove that $\mathbb{P}(E^c) \leq nK\delta$.

35.2 Improve the bound in Theorem 35.1 to show that $\text{BR}_n \leq C\sqrt{Kn}$ where $C > 0$ is a universal constant.



Replace the naive confidence intervals used in the proof of Theorem 35.1 by the more refined confidence bounds used in Chapter 9. The source for this result is the paper by [Bubeck and Liu \[2013\]](#).

35.3 Prove the inequality in Eq. (35.5).

35.4 Prove Lemmas 35.1 and 35.2.

35.5 Suppose that X and Y are independent random variables. Show that $I(X; Y) = 0$.

35.6 Let \mathcal{E} be a set of bandits and \mathbb{Q} a prior on \mathcal{E} .

- (a) Recall that $R_n^*(\mathcal{E}) = \inf_{\pi} \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu)$ is the minimax regret. Show that $R_n^*(\mathcal{E}) \geq \inf_{\pi} \text{BR}_n(\mathcal{E}, \mathbb{Q})$.
- (b) Let \mathcal{E} be the set of Bernoulli bandits. Find a sequence of priors (\mathbb{Q}_n) such that $\text{BR}_n(\mathcal{E}, \mathbb{Q}_n) \geq c\sqrt{Kn}$ for all $n \geq K$ where $c > 0$ is a universal constant.

35.7 Prove that for any prior such that $X_{ti} \in [0, 1]$ almost surely the Bayesian regret of information-directed sampling satisfies

$$\text{BR}_n \leq \sqrt{\frac{Kn \log(K)}{2}}.$$

35.8 The purpose of this exercise is to compare Thompson sampling for Gaussian bandits with UCB.

- (a) Implement the Gaussian Thompson sampling algorithm described by Eq. (35.3).
- (b) Compare the expected regret of Thompson sampling with the version of UCB in Chapter 8 and refinements in Eq. (9.2) and Eq. (9.3).
- (c) What about the variance of these algorithms?
- (d) Briefly explain the pros and cons of Thompson sampling relative to UCB.

35.9 Fix a Gaussian bandit with unit variance and mean vector $\mu = (0, 1/10)$ and horizon $n = 1000$. Now consider Thompson sampling with a Gaussian model with known unit covariance and a prior on the unknown mean of each arm given by a Gaussian distribution with mean μ_P and covariance $\sigma_P^2 I$.

- (a) Let the prior mean be $\mu_P = (0, 0)$ and plot the regret of Thompson sampling as a function of the prior variance σ_P^2 .
- (b) Repeat the above with $\mu_P = (0, 1/10)$ and $(0, -1/10)$ and $(2/10, 1/10)$.
- (c) Explain your results.

Part VIII

Beyond Bandits

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

36 Partial Monitoring

While in a bandit problem the feedback that the learner receives from the environment is the loss (or reward) of the chosen action, in **partial monitoring** the coupling between the loss of the action and the feedback received by the learner is loosened.

To illustrate the ideas we consider the problem of learning to match pennies when feedback is costly. Let $c > 0$ be a known constant. At the start of the game the adversary secretly chooses a sequence $i_1, \dots, i_n \in \{\text{heads, tails}\}$. In each round the learner chooses action $A_t \in \{\text{heads, tails, uncertain}\}$ and the loss for choosing action a in round t is

$$y_{ta} = \begin{cases} 0, & \text{if } a = i_t; \\ c, & \text{if } a = \text{uncertain}; \\ 1, & \text{otherwise.} \end{cases}$$

So far this looks like a bandit problem. The difference is that the learner never directly observes y_{tA_t} . Instead, the learner observes nothing unless $A_t = \text{uncertain}$ in which case they observe the value of i_t . As usual, the goal of the regret is to minimize the regret, which is

$$R_n = \mathbb{E} \left[\max_{a \in [K]} \sum_{t=1}^n (y_{tA_t} - y_{ta}) \right].$$

How should a learner act in problems like this, where the loss is not directly observed? Can we find a policy with sublinear regret? In this chapter we give nearly complete answers to these questions for a large class of finite adversarial partial monitoring problems.



Matching pennies with costly feedback seems like an esoteric problem. But think about adding contextual information and replace the pennies with emails to be classified as spam or otherwise. The true label is only accessible by asking a human, which replaces the third action.

36.1 Finite adversarial partial monitoring problems

To reduce clutter we slightly abuse notation by using (e_i) to denote the standard basis vectors of Euclidean spaces of potentially different dimensions. A K -action, E -outcome, F -feedback finite adversarial partial monitoring problem is specified by a **loss matrix** $\mathcal{L} \in \mathbb{R}^{K \times E}$ and a **feedback matrix** $\Phi \in [F]^{K \times E}$. At the beginning of the game, the learner gets \mathcal{L} and Φ , while the environment secretly chooses n outcomes i_1, \dots, i_n with $i_t \in [E]$. The loss of action $a \in [K]$ in round t is $y_{ta} = \mathcal{L}_{ai_t}$. In each round t the learner chooses $A_t \in [K]$ and receives feedback $\Phi_t = \Phi_{A_t i_t}$. Given partial monitoring problem $G = (\Phi, \mathcal{L})$ the regret of policy π in environment $i_{1:n} = (i_t)_{t=1}^n$ is

$$R_n(\pi, i_{1:n}, G) = \max_{a \in [K]} \mathbb{E}_{\pi, i_{1:n}, G} \left[\sum_{t=1}^n (y_{tA_t} - y_{ta}) \right].$$

The index of the expectation operator is a reminder that the distribution of $\{y_{tA_t}\}$ is dependent on π , $i_{1:n}$ and G . We will omit these indices and the arguments of R_n when they can be inferred from the context.

36.1.1 Examples

The partial monitoring framework is rich enough to model a wide variety of problems, a few of which are illustrated by the examples that follow. Many of the examples are not very interesting on their own, but are included to highlight the flexibility of the framework and challenges of making the regret small.

EXAMPLE 36.1 (Hopeless problem) Some partial monitoring problems are completely hopeless in the sense one cannot expect to make the regret small. A simple example occurs when $K = E = 2$ and $F = 1$ and

$$\mathcal{L} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \Phi = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \quad (36.1)$$

Note that rows/columns correspond to choices of the learner/environment respectively. In both rows (corresponding to the actions of the learner), the feedback matrix has identical entries for both columns. As the learner has no way of distinguishing between different sequences of outcomes, there is no way to learn and avoid linear regret.



Two feedback matrices $\Phi, \Phi' \in [F]^{K \times E}$ encode the same information if the pattern of identical entries in each row match. More precisely, if for each row $a \in [K]$ there is an injective function $\sigma : [E] \rightarrow [E]$ such that $\Phi'_{ai} = \sigma(\Phi_{ai})$ for all $i \in [E]$.

EXAMPLE 36.2 (Trivial problem) Just as there are hopeless problems, there

are also trivial problems. For example, when one action dominates all others as in the following problem:

$$\mathcal{L} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \quad \Phi = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Clearly, in this game the learner can safely ignore the second action and suffer zero regret, regardless of the choices of the adversary.

EXAMPLE 36.3 (Matching pennies) The penny-matching problem mentioned in the introduction has $K = 3$ actions $E = 2$ outcomes and is described by

$$\mathcal{L} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ c & c \end{pmatrix}, \quad \Phi = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}. \quad (36.2)$$

Matching pennies is a hard game for $c > 1/2$ in the sense that the adversary can force the regret of any adversary to be at least $\Omega(n^{2/3})$. To see this, consider the randomized adversary that chooses the first outcome with probability p and the second with probability $1 - p$. Let $\varepsilon > 0$ be a small constant to be chosen later and assume p is either $1/2 + \varepsilon$ or $1/2 - \varepsilon$, which determines two environments. The techniques in Chapter 13 show that the learner can only distinguish between these environments by playing the third action about $1/\varepsilon^2$ times. If the learner does not choose to do this, then the regret is expected to be $\Omega(n\varepsilon)$. Taking these together shows the regret is lower bounded by $R_n = \Omega(\min(n\varepsilon, (c - 1/2 + \varepsilon)/\varepsilon^2))$. Choosing $\varepsilon = n^{-1/3}$ leads to a bound of $R_n = \Omega((c - 1/2)n^{2/3})$. Notice the argument fails when $c \leq 1/2$. We encourage you to pause for a minute to convince yourself about the correctness of the above argument and to consider what might be the situation when $c \leq 1/2$.

EXAMPLE 36.4 (Bandits) Finite-armed adversarial bandits with binary losses can be represented in the partial monitoring framework. When $K = 2$ this is possible with the following matrices:

$$\mathcal{L} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad \Phi = \begin{pmatrix} 1 & 2 & 1 & 2 \\ 1 & 1 & 2 & 2 \end{pmatrix}.$$

The number of columns for this game is 2^K . For non-binary rewards you would need even more columns. A partial monitoring problem where $\Phi = \mathcal{L}$ can be called a bandit problem because the learner observes the loss of the chosen action. In bandit games we can simply use Exp3 to guarantee a regret of $O(\sqrt{Kn})$.

EXAMPLE 36.5 (Full information problems) One can also represent problems where the learner observes all the losses. With binary losses and two actions we have

$$\mathcal{L} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad \Phi = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}.$$

Like for bandits the size of the game grows very quickly as more actions/outcomes are added.

EXAMPLE 36.6 (Dynamic pricing) A charity worker is going door-to-door selling calendars. The marginal cost of a calendar is close to zero, but the wages of the door-knocker represents a fixed cost of $c > 0$ per occupied house. The question is how to price the calendar. Each round corresponds to an attempt to sell a calendar and the action is the seller's asking price from one of E choices. The potential buyer will purchase the calendar if the asking price is low enough. Below we give the corresponding matrices for case where both the candidate asking prices and the possible values for the buyer's private valuations are $\{\$1, \$2, \$3, \$4\}$:

$$\mathcal{L} = \begin{pmatrix} c & c-1 & c-1 & c-1 & c-1 \\ c & c & c-2 & c-2 & c-2 \\ c & c & c & c-3 & c-3 \\ c & c & c & c & c-4 \end{pmatrix}, \quad \Phi = \begin{pmatrix} 1 & 2 & 2 & 2 & 2 \\ 1 & 1 & 2 & 2 & 2 \\ 1 & 1 & 1 & 2 & 2 \\ 1 & 1 & 1 & 1 & 2 \end{pmatrix}.$$

Notice that observing the feedback is sufficient to deduce the loss so the problem could be tackled with a bandit algorithm. But there is additional structure in the losses here because the learner knows that if a calendar did not sell for \$3 then it would not sell for \$4.

36.2 The structure of partial monitoring

The minimax regret of partial monitoring problem $G = (\mathcal{L}, \Phi)$ is

$$R_n^*(G) = \inf_{\pi} \max_{u_{1:n}} R_n(\pi, u_{1:n}, G).$$

One of the core questions in partial monitoring is to understand the growth of $R_n^*(G)$ as a function of n for different games. We have seen examples where

$$\begin{aligned} R_n^*(G) &= 0 && \text{(Example 36.2)} \\ R_n^*(G) &= \tilde{\Theta}(\sqrt{n}) && \text{(Example 36.4)} \\ R_n^*(G) &= \Theta(n^{2/3}) && \text{(Example 36.3)} \\ R_n^*(G) &= \Omega(n). && \text{(Example 36.1)} \end{aligned}$$

The main result of this chapter is that there are no other options. A partial monitoring problem is called **trivial** if $R_n^*(G) = 0$, **easy** if $R_n^*(G) = \tilde{\Theta}(\sqrt{n})$, **hard** if $R_n^*(G) = \Theta(n^{2/3})$ and **hopeless** if $R_n^*(G) = \Omega(n)$. Furthermore, we will show that the category of any G can be deduced from elementary linear algebra.

What makes matching pennies hard and bandits easy? To get a handle on this we need a geometric representation of partial monitoring problems. The next few paragraphs introduce a lot of new terminology that can be hard to grasp all at once. At the end of the section there is an example illustrating the concepts.

The geometry underlying partially monitoring comes from viewing the problem as a linear prediction problem, where both the adversary and the learner play on some simplex. Starting with reworking the adversary's choices, let $u_t = e_{i_t} \in \mathbb{R}^E$, where e_1, \dots, e_E are the standard basis vectors. Then, we can equivalently think of the environment choosing the sequence $\{u_t\}_{t=1}^n$. Letting $\ell_a \in \mathbb{R}^E$ be the a th row of matrix \mathcal{L} , where $a \in [K]$, we have that $y_{ta} = \langle \ell_a, u_t \rangle$ is the loss suffered when choosing action a in round t .

Let $\bar{u}_t = \frac{1}{t} \sum_{s=1}^t u_s \in \mathcal{P}_{E-1}$ be the vector of mean frequencies of the adversary's choices over t rounds. An action a is optimal in hindsight if $\max_b \langle \ell_a - \ell_b, \bar{u}_n \rangle = 0$. The **cell** of an action a is subset of \mathcal{P}_{E-1} on which it is optimal:

$$C_a = \left\{ u \in \mathcal{P}_{E-1} : \max_{b \in [K]} \langle \ell_a - \ell_b, u \rangle \leq 0 \right\},$$

which is convex polytope. The collection $\{C_a : a \in [K]\}$ is called the **cell decomposition**. Actions with $C_a = \emptyset$ are called **dominated** because they are never optimal, no matter how the adversary plays. For nondominated actions we define the **dimension** of an action to be the dimension of the **affine hull** of C_a . Readers unfamiliar with the affine hull should read Note 3 at the end of the chapter. A nondominated action is called **Pareto optimal** if it has dimension $E - 1$ and **degenerate** otherwise. Actions a and b are **duplicates** if $\ell_a = \ell_b$. Pareto optimal actions a and b are **neighbors** if $C_a \cap C_b$ has dimension $E - 2$. Note that if a and b are Pareto optimal duplicates, then $C_a \cap C_b$ has dimension $E - 1$ and the definition means that a and b are not neighbors. For Pareto optimal action a we let \mathcal{N}_a be the set consisting of a and its neighbors. Given a pair of neighbors (a, b) we let $\mathcal{N}_{ab} = \{c \in [K] : C_a \cap C_b \subseteq C_c\}$, while for Pareto optimal action a we let $\mathcal{N}_{aa} = \emptyset$.



Dominated and degenerate actions can never be uniquely optimal in hindsight, but their presence can make the difference between a hard game and a hopeless one. If $c > 1/2$, then the third action in matching pennies is dominated, but without it the learner would suffer linear regret. Duplicate actions are only duplicate in the sense that they have the same loss. They may have different feedback structures and so cannot be trivially combined.

Let a and b be neighboring actions. The next lemma characterizes actions in \mathcal{N}_{ab} as either a, b , duplicates of a, b or degenerate actions d for which ℓ_d is a convex combination of ℓ_a and ℓ_b . The situation is illustrated when $E = 2$ in Fig. 36.1.

LEMMA 36.1 *Let a, b be neighboring actions and $d \in \mathcal{N}_{ab}$ be an action such that $\ell_d \notin \{\ell_a, \ell_b\}$. Then*

- (a) *There exists an $\alpha \in (0, 1)$ such that $\ell_d = \alpha \ell_a + (1 - \alpha) \ell_b$.*
- (b) *$C_d = C_a \cap C_b$.*

(c) d has dimension $E - 2$.

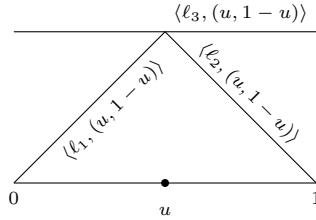


Figure 36.1 The figure shows the situation when $E = 2$ and $\ell_1 = (1, 0)$ and $\ell_2 = (0, 1)$ and $\ell_3 = (1/2, 1/2)$. Then $C_1 = [0, 1/2]$ and $C_2 = [1/2, 1]$, which both have dimension $1 = E - 1$. Then $C_3 = \{1/2\} = C_1 \cap C_2$, which has dimension 0.

Proof We use the fact that if $\mathcal{X} \subseteq \mathcal{Y} \subseteq \mathbb{R}^d$ and $\dim(\mathcal{X}) = \dim(\mathcal{Y})$, then $\text{aff}(\mathcal{X}) = \text{aff}(\mathcal{Y})$ (Exercise 36.1). Clearly $C_a \cap C_b \subseteq C_a \cap C_d$ and $\text{aff}(C_a \cap C_b) = \ker(\ell_a - \ell_b)$ and $\text{aff}(C_a \cap C_d) = \ker(\ell_a - \ell_d)$. By assumption $\dim(C_a \cap C_b) = E - 2$. Since $C_a \cap C_b \subseteq C_a \cap C_d$ it holds that $\dim(C_a \cap C_d) \geq E - 2$. Furthermore, $\dim(C_a \cap C_d) \leq E - 2$, since otherwise $\ell_d = \ell_a$. Hence $\ker(\ell_a - \ell_b) = \ker(\ell_a - \ell_d)$, which means that $\ell_a - \ell_b$ is proportional to $\ell_a - \ell_d$ so that $(1-\alpha)(\ell_a - \ell_b) = \ell_a - \ell_d$ for some $\alpha \neq 1$. Rearranging shows that

$$\ell_d = \alpha \ell_a + (1 - \alpha) \ell_b.$$

Now we show that $\alpha \in (0, 1)$. First note that $\alpha \notin \{0, 1\}$ since otherwise $\ell_d \in \{\ell_a, \ell_b\}$. Let $u \in C_a$ be such that $\langle \ell_a, u \rangle < \langle \ell_b, u \rangle$, which exists since $\dim(C_a) = E - 1$ and $\dim(C_a \cap C_b) = E - 2$. Then

$$\langle \ell_a, u \rangle \leq \langle \ell_d, u \rangle = \alpha \langle \ell_a, u \rangle + (1 - \alpha) \langle \ell_b, u \rangle = \langle \ell_a, u \rangle + (\alpha - 1) \langle \ell_a - \ell_b, u \rangle,$$

which by the positivity of $\langle \ell_a - \ell_b, u \rangle$ implies that $\alpha \leq 1$. A symmetric argument shows that $\alpha > 0$. For (b), it suffices to show that $C_d \subset C_a \cap C_b$. By de Morgan's law for this it suffices to show that $\mathcal{P}_{E-1} \setminus (C_a \cap C_b) \subset \mathcal{P}_{E-1} \setminus C_d$. Thus, pick some $u \in \mathcal{P}_{E-1} \setminus (C_a \cap C_b)$. The goal is to show that $u \notin C_d$. The choice of u implies that there exists an action c such that $\langle \ell_a - \ell_c, u \rangle \geq 0$ and $\langle \ell_b - \ell_c, u \rangle \geq 0$ with a strict inequality for either a or b (or both). Therefore using the fact that $\alpha \in (0, 1)$ we have

$$\langle \ell_d, u \rangle = \alpha \langle \ell_a, u \rangle + (1 - \alpha) \langle \ell_b, u \rangle > \langle \ell_c, u \rangle,$$

which by definition means that $u \notin C_d$, completing the proof of (b). Finally, (c) is immediate from (b) and the definition of neighboring actions. \square

In order to achieve small regret the learner needs to identify an optimal action. How efficiently this can be done depends on the feedback matrix. First, note that given access to the loss matrix, the learner can restrict the search for the optimal action to the Pareto optimal actions. One way to find the optimal action then

could be to estimate $\langle \ell_a, u_t \rangle$ for each Pareto optimal action a and $t \in [n]$ and take differences of the estimates to compare actions. This is asking too much, and a better option is to estimate $\langle \ell_a - \ell_b, u_t \rangle$ directly. This is a better option because on the one hand it is clearly necessary to know the loss differences between Pareto optimal actions, and on the other hand there exist games for which $\langle \ell_a, u_t \rangle$ cannot be estimated, but the differences can. For example, the following game has this property.

$$\mathcal{L} = \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \end{pmatrix}, \quad \Phi = \begin{pmatrix} 1 & 2 & 1 & 2 \\ 2 & 1 & 2 & 1 \end{pmatrix}$$

The learner can never tell if the environment is playing in the first two columns or the last two, but the differences between the losses are easily deduced from the feedback. We emphasize once again that only the loss differences between Pareto optimal actions need to be estimated: There are in fact games that are easy yet some loss differences cannot be estimated. For example, there is never any need to estimate the losses of a dominated action.

Having decided we need to estimate the loss differences for Pareto optimal actions, the next question is how can the learner do this? Suppose in round t the learner samples A_t from distribution $P_t \in \text{ri}(\mathcal{P}_{K-1})$. Let a and b be Pareto optimal and suppose we want an estimator $\hat{\Delta}$ of $\Delta = y_{ta} - y_{tb}$. Our estimator $\hat{\Delta}$ should depend on A_t and Φ_t , which suggests defining

$$\hat{\Delta} = \frac{v(A_t, \Phi_t)}{P_{tA_t}},$$

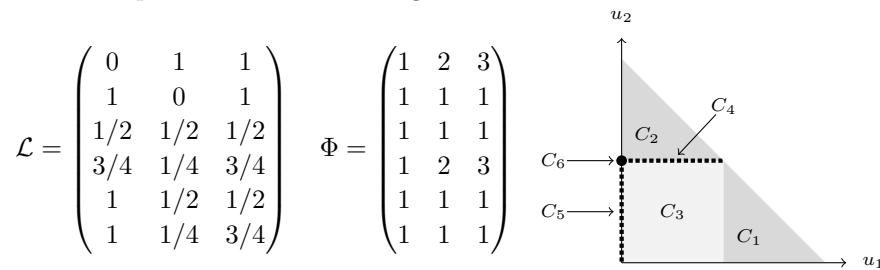
where $v : [K] \times [F] \rightarrow \mathbb{R}$ is some suitable chosen function. The division by P_{tA_t} is a convenient normalization and could be pushed into v . The reader can check that $\hat{\Delta}$ is unbiased regardless the choice u_t if and only if

$$\ell_{ai} - \ell_{bi} = \sum_{c=1}^K v(c, \Phi_{ci}) \quad \text{for all } i \in [E]. \quad (36.3)$$

A pair of Pareto optimal actions a and b are called **globally observable** if there exists a function v satisfying Eq. (36.3). They are **locally observable** if the function can be chosen so that $v(c, f) = 0$ whenever $c \notin \mathcal{N}_{ab}$. A partial monitoring problem $G = (\mathcal{L}, \Phi)$ is called globally/locally observable if all pairs of neighboring actions are globally/locally observable. The global/local observability conditions formalize the idea introduced in Example 36.3. Games that are globally observable but not locally observable are hard because the learner cannot identify the optimal action by playing near-optimal actions only. Instead it has to play badly suboptimal actions to gain information and this increases the minimax regret.

EXAMPLE 36.7 The partial monitoring problem below has six actions, three feedbacks and three outcomes. The cell decomposition is shown on the right with the 2-simplex parameterized by its first two coordinates u_1 and u_2 so that

$u_3 = 1 - u_2 - u_1$. Actions 1, 2 and 3 are Pareto optimal. There are no dominated actions while actions 4 and 5 are 1-dimensional and action 6 is 0-dimensional. The neighbors are $(1, 3)$ and $(2, 3)$, which are both locally observable and so the game is locally observable. Note that $(1, 2)$ are *not* neighbors because the intersection of their cells is $(E - 3)$ -dimensional. Finally, $\mathcal{N}_3 = \{1, 2, 3\}$ and $\mathcal{N}_1 = \{1, 3\}$ and $\mathcal{N}_{23} = \{2, 3, 4\}$. Think about how we decided on what losses to use to get the cell decomposition shown in the figure!



36.3 Classification of finite adversarial partial monitoring

The terminology in the last chapter finally allows us to state the main theorem of this chapter that classifies finite adversarial partial monitoring games.

THEOREM 36.1 *The minimax regret of partial monitoring problem $G = (\mathcal{L}, \Phi)$ falls into one of four categories:*

$$R_n^*(G) = \begin{cases} 0, & \text{if } G \text{ has no pairs of neighboring actions;} \\ \tilde{\Theta}(\sqrt{n}), & \text{if } G \text{ is locally observable and has neighboring actions;} \\ \Theta(n^{2/3}), & \text{if } G \text{ is globally observable, but not locally observable;} \\ \Omega(n), & \text{otherwise.} \end{cases}$$



The Landou notation is used in the traditional mathematical sense and obscures dependence on K, E, F and the finer structure of $G = (\mathcal{L}, \Phi)$.

The proof is split into parts by proving upper and lower bounds for each part. First up is the lower bounds. We then describe a policy for locally observable games and analyze its regret. The upper bound for globally observable games is left as an exercise to the reader (Exercise 36.11).

36.4 Lower bounds

Like for bandits, the lower bounds are most easily proven using a stochastic adversary. In stochastic partial monitoring we assume that u_1, \dots, u_n are chosen at independently at random from the same distribution. To emphasize the

randomness we switch to capital letters. Given partial monitoring problem $G = (\mathcal{L}, \Phi)$ and probability vector $u \in \mathcal{P}_{E-1}$ the stochastic partial monitoring environment associated with u samples a sequence of independently and identically distribution random variables I_1, \dots, I_n with $\mathbb{P}(I_t = i) = u_i$ and $U_t = e_{I_t}$. In each round t a policy chooses action A_t and receives feedback $\Phi_t = \Phi_{A_t I_t}$. The regret is

$$R_n(\pi, u, G) = \max_{a \in [K]} \mathbb{E} \left[\sum_{t=1}^n \langle \ell_{A_t} - \ell_a, U_t \rangle \right] = \max_{a \in [K]} \mathbb{E} \left[\sum_{t=1}^n \langle \ell_{A_t} - \ell_a, u \rangle \right].$$

The reader should check that $R_n^*(G) \geq \min_{\pi} \max_{u \in \mathcal{P}_{E-1}} R_n(\pi, u, G)$, which allows us to restrict our attention to stochastic partial monitoring problems. Given $u, q \in \mathcal{P}_{E-1}$, let $D(u, q)$ be the relative entropy between categorical distributions with parameters u and q respectively:

$$D(u, q) = \sum_{i=1}^K u_i \log \left(\frac{u_i}{q_i} \right) \leq \sum_{i=1}^K \frac{(u_i - q_i)^2}{q_i}, \quad (36.4)$$

where the second inequality follows from the fact that for measures P, Q we have $D(P, Q) \leq \chi^2(P, Q)$ (see Note 4 in Chapter 13).

THEOREM 36.2 *Let $G = (\mathcal{L}, \Phi)$ be a globally observable partial monitoring problem that is not locally observable. Then there exists a constant $c_G > 0$ such that $R_n^*(G) \geq c_G n^{2/3}$.*

Proof The proof involves several steps. Roughly, we need to define two alternative stochastic partial monitoring problems. We then show these environments are hard to distinguish without playing an action associated with a large loss. Finally we balance the cost of distinguishing the environments against the linear cost of playing randomly.

Step 1: Defining the alternatives

Let a, b be a pair neighboring actions that are not locally observable. Then by definition $C_a \cap C_b$ is a polytope of dimension $E - 2$. Let u be the centroid of $C_a \cap C_b$ and

$$\varepsilon = \min_{c \notin \mathcal{N}_{ab}} \langle \ell_c - \ell_a, u \rangle. \quad (36.5)$$

The value of ε is well defined, since by global observability of G , but nonlocal observability of (a, b) there must exist some action $c \notin \mathcal{N}_{ab}$. Furthermore, since $c \notin \mathcal{N}_{ab}$ it follows that $\varepsilon > 0$. As in the lower bound constructions for stochastic bandits, we now define two stochastic partial monitoring problems. Since (a, b) are not locally observable, there does not exist a function $v : [K] \times [F] \rightarrow \mathbb{R}$ such that for all $i \in [E]$,

$$\sum_{c \in N_k} v(c, \Phi_{ci}) = \ell_{ai} - \ell_{bi}. \quad (36.6)$$

In this form it does not seem obvious what the next step should be. To clear things up a little we introduce some linear algebra. Let $S_c \in \{0, 1\}^{F \times E}$ be the matrix with $(S_c)_{fi} = \mathbb{I}\{\Phi_{ci} = f\}$, which is chosen so that $S_c e_i = e_{\Phi_{ci}}$. Define the linear map $S : \mathbb{R}^E \rightarrow \mathbb{R}^{|\mathcal{N}_{ab}|F}$ by

$$S = \begin{pmatrix} S_a \\ S_b \\ \vdots \\ S_c \end{pmatrix},$$

which is the matrix formed by stacking the matrices $\{S_c : c \in \mathcal{N}_{ab}\}$. Then there exists a v satisfying Eq. (36.6) if and only if there exists a $w \in \mathbb{R}^{|\mathcal{N}_{ab}|F}$ such that

$$(\ell_a - \ell_b)^\top = w^\top S.$$

In other words, actions (a, b) are locally observable if and only if $\ell_a - \ell_b \in \text{im}(S^\top)$. Since we have assumed that (a, b) are not locally observable, it means that $\ell_a - \ell_b \notin \text{im}(S^\top)$. Let $z \in \text{im}(S^\top)$ and $w \in \ker(S)$ be such that $\ell_a - \ell_b = z + w$, which is possible since $\text{im}(S^\top) \oplus \ker(S) = \mathbb{R}^E$. Since $\ell_a - \ell_b \notin \text{im}(S^\top)$ it holds that $w \neq 0$ and $\langle \ell_a - \ell_b, w \rangle = \langle z + w, w \rangle = \langle w, w \rangle \neq 0$. Finally let $q = w / \langle \ell_a - \ell_b, w \rangle$. To summarize, we have demonstrated the existence of a vector $q \in \mathbb{R}^E$, $q \neq 0$ such that $Sq = 0$ and $\langle \ell_a - \ell_b, q \rangle = 1$. Let $\Delta > 0$ be some small constant to be tuned subsequently and define $u_a = u - \Delta q$ and $u_b = u + \Delta q$ so that

$$\langle \ell_b - \ell_a, u_a \rangle = \Delta \quad \text{and} \quad \langle \ell_a - \ell_b, u_b \rangle = \Delta.$$

We note that if Δ is sufficiently small, then $u_a \in C_a$ and $u_b \in C_b$. This means that action a is optimal if the environment plays u_a on average and b is optimal if the environment plays u_b on average (see Fig. 36.2).

Step 2: Calculating the relative entropy

Given action c and $w \in \mathcal{P}_{E-1}$ let \mathbb{P}_{cw} be the distribution on the feedback observed by the learner when playing action c in stochastic partial monitoring environment determined by w . That is $\mathbb{P}_{cw}(f) = \mathbb{P}_w(\Phi_t = f | A_t = c) = (S_c w)_f$. Further, let \mathbb{P}_w be the distribution on the histories $H_n = (A_1, \Phi_1, \dots, A_n, \Phi_n)$ arising from the interaction of the learner's policy with the stochastic environment determined by w . A modification of Lemma 15.1 shows that

$$D(\mathbb{P}_{u_a}, \mathbb{P}_{u_b}) = \sum_{c \in [K]} \mathbb{E}[T_c(n)] D(\mathbb{P}_{cu_a}, \mathbb{P}_{cu_b}), \quad (36.7)$$

By the definitions of u_a and u_b , we have $S_c u_a = S_c u_b$ for all $c \in \mathcal{N}_{ab}$. Therefore $\mathbb{P}_{cu_a} = \mathbb{P}_{cu_b}$ and so $D(\mathbb{P}_{cu_a}, \mathbb{P}_{cu_b}) = 0$ for all $c \in \mathcal{N}_{ab}$. On the other hand, if $c \notin \mathcal{N}_{ab}$, then by Eq. (36.4),

$$D(\mathbb{P}_{cu_a}, \mathbb{P}_{cu_b}) \leq D(u_a, u_b) \leq \sum_{i=1}^E \frac{(u_{ai} - u_{bi})^2}{u_{bi}} = 4\Delta^2 \sum_{i=1}^K \frac{q_i^2}{u_i + \Delta q_i} \leq C_u \Delta^2,$$

where C_u is a suitably large constant. We note that $u \in C_a \cap C_b$ is not on the boundary of \mathcal{P}_{E-1} , so $u_i > 0$ for all i . Therefore

$$D(\mathbb{P}_{u_a}, \mathbb{P}_{u_b}) \leq c_U \sum_{c \in [K]} \mathbb{E}[T_c(n)] \Delta^2. \quad (36.8)$$

Step 3: Comparing the regret

By Eq. (36.5) and Hölder's inequality, for $c \notin \mathcal{N}_{ab}$ we have $\langle \ell_c - \ell_a, u_a \rangle = \varepsilon + \langle \ell_c - \ell_a, \Delta q \rangle \geq \varepsilon - \Delta \|q\|_1$ and $\langle \ell_c - \ell_b, u_b \rangle \geq \varepsilon - \Delta \|q\|_1$, where, for simplicity, and without the loss of generality, we assumed that the losses lie in $[0, 1]$. Define $\tilde{T}(n)$ to be the number of times an arm not in \mathcal{N}_{ab} is played:

$$\tilde{T}(n) = \sum_{c \notin \mathcal{N}_{ab}} T_c(n).$$

By Lemma 36.1, for each action $c \in \mathcal{N}_{ab}$ there exists an $\alpha \in [0, 1]$ such that $\ell_c = \alpha \ell_a + (1 - \alpha) \ell_b$. Therefore

$$\langle \ell_c - \ell_a, u_a \rangle + \langle \ell_c - \ell_b, u_b \rangle = (1 - \alpha) \langle \ell_b - \ell_a, u_a \rangle + \alpha \langle \ell_a - \ell_b, u_b \rangle = \Delta, \quad (36.9)$$

which means that $\max(\langle \ell_c - \ell_a, u_a \rangle, \langle \ell_c - \ell_b, u_b \rangle) \geq \Delta/2$. Define $\bar{T}(n)$ as the number of times an arm in \mathcal{N}_{ab} is played that is at least $\Delta/2$ suboptimal in u_a :

$$\bar{T}(n) = \sum_{c \in \mathcal{N}_{ab}} \mathbb{I} \left\{ \langle \ell_c - \ell_a, u_a \rangle \geq \frac{\Delta}{2} \right\} T_c(n).$$

It also follows from (36.9) that if $c \in \mathcal{N}_{ab}$ and $\langle \ell_c - \ell_a, u_a \rangle < \frac{\Delta}{2}$ then $\langle \ell_c - \ell_b, u_b \rangle \geq \frac{\Delta}{2}$. Hence, under u_b the random pseudo-regret, $\sum_c T_c(n) \langle \ell_c - \ell_b, u_b \rangle$, is at least $(n - \bar{T}(n))\Delta/2$. Assume that Δ is chosen sufficiently small so that $\Delta \|q\|_1 \leq \varepsilon/2$. Then, by the above,

$$\begin{aligned} R_n(\pi, u_a, G) + R_n(\pi, u_b, G) \\ &= \mathbb{E}_{u_a} \left[\sum_{c \in [K]} T_c(n) \langle \ell_c - \ell_a, u_a \rangle \right] + \mathbb{E}_{u_b} \left[\sum_{c \in [K]} T_c(n) \langle \ell_c - \ell_b, u_b \rangle \right] \\ &\geq \frac{\varepsilon}{2} \mathbb{E}_{u_a} [\tilde{T}(n)] + \frac{n\Delta}{4} (\mathbb{P}_{u_a}(\bar{T}(n) \geq n/2) + \mathbb{P}_{u_b}(\bar{T}(n) < n/2)) \\ &\geq \frac{\varepsilon}{2} \mathbb{E}_{u_a} [\tilde{T}(n)] + \frac{n\Delta}{8} \exp(-D(\mathbb{P}_{u_a}, \mathbb{P}_{u_b})) \\ &\geq \frac{\varepsilon}{2} \mathbb{E}_{u_a} [\tilde{T}(n)] + \frac{n\Delta}{8} \exp(-C_u \Delta^2 \mathbb{E}_{u_a} [\tilde{T}(n)]), \end{aligned}$$

where the second inequality follows from Theorem 14.2 and the third from Eqs. (36.7) and (36.8). The bound is completed by choosing $\Delta = \varepsilon/(2\|q\|_1 n^{1/3})$ (which is finite since $q \neq 0$) and straightforward optimization (Exercise 36.5). \square

We leave the following theorems as exercises for the reader (Exercises 36.6 and 36.7).

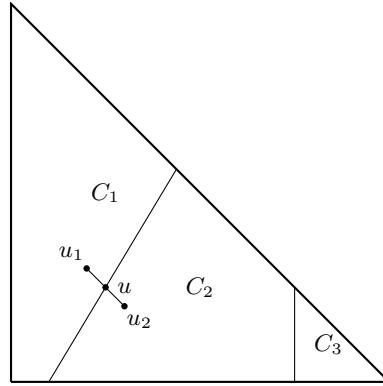


Figure 36.2 Lower bound construction for hard partial monitoring problems

THEOREM 36.3 *If G is not globally observable and has at least two non-dominated actions, then there exists a constant $c_G > 0$ such that $R_n^*(G) \geq c_G n$.*

Proof sketch Since G is not globally observable there exists a pair of neighboring actions (a, b) that are not globally observable. Let u be the centroid of $C_a \cap C_b$. Let $S \in \mathbb{R}^{KF \times E}$ be the stack of matrices from $\{S_c : c \in [K]\}$ (all actions). Then using the same argument as the previous proof we have $\ell_a - \ell_b \notin \text{im}(S^\top)$. Now define $q \in \mathbb{R}^E$ such that $\langle \ell_a - \ell_b, q \rangle = 1$ and $Sq = 0$. Let $\Delta > 0$ be sufficiently small and $u_a = u - \Delta q$ and $u_b = u + \Delta q$. Show that $D(\mathbb{P}_{u_a}, \mathbb{P}_{u_b}) = 0$ for all policies and complete the proof in the same fashion as the proof of Theorem 36.2. \square

THEOREM 36.4 *Let $G = (\mathcal{L}, \Phi)$ be locally observable and have at least one pair of neighbours. Then there exists a constant $c_G > 0$ such that for all large enough n the minimax regret satisfies $R_n^*(G) \geq c_G \sqrt{n}$.*

Proof sketch By assumption there exists a pair of neighbouring actions (a, b) . Define u as the centroid of $C_a \cap C_b$ and let $q = (\ell_a - \ell_b)/\|\ell_a - \ell_b\|^2$. For sufficiently small $\Delta > 0$ let $u_a = u - \Delta q$ and $u_b = u + \Delta q$. Then

$$D(\mathbb{P}_{u_a}, \mathbb{P}_{u_b}) \leq n \sum_{i=1}^E \frac{(u_{ai} - u_{bi})^2}{u_{bi}} \leq C_G n \Delta^2,$$

where $C_G > 0$ is a game-dependent constant. Let $\Delta = 1/\sqrt{n}$ and apply the ideas in the proof of Theorem 36.2. \square

36.5 Policy for easy games

Fix a locally observable game $G = (\mathcal{L}, \Phi)$ with at least one pair of neighboring actions. We describe a policy called NeighborhoodWatch2. In every round the policy always chooses $A_t \in \cup_{a,b} \mathcal{N}_{ab}$ where the union is over pairs of neighboring actions. For example, in the partial monitoring game described

in Example 36.7 the policy would only play actions 1, 2, 3 and 4. Removing degenerate actions can only increase the minimax regret, so from now on we assume that $[K] = \cup_{a,b} \text{neighbors}(\mathcal{N}_{ab})$. We let \mathcal{A} be an arbitrary largest subset of Pareto optimal actions such that \mathcal{A} does not contain actions that are duplicates of each other and $\mathcal{D} = [K] \setminus \mathcal{A}$ be the remaining actions. In each round t the policy performs four steps as described below.

Step 1 (Local games)

For each $k \in \mathcal{A}$ the policy maintains an exponential weights distribution over $\mathcal{A} \cup \mathcal{D} = [K]$, but concentrated on the intersection of the neighborhood \mathcal{N}_k of k and \mathcal{A} (recall that \mathcal{N}_k contains the neighbors of k , some of which may be duplicates of each other). We denote this distribution by $Q_{tka} \in \mathcal{P}_{K-1}$. For $a \in [K]$, the value of Q_{tka} is given

$$Q_{tka} = \frac{\mathbb{I}_{\mathcal{N}_k \cap \mathcal{A}}(a) \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{ska}\right)}{\sum_{b \in \mathcal{N}_k \cap \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{skb}\right)},$$

where $\eta > 0$ is the learning rate and the \tilde{Z}_{ska} are estimators of the loss difference $y_{sa} - y_{sk}$ and will be introduced in step four below. For actions $k \in \mathcal{D}$ we define $Q_{tka} = \mathbb{I}_{\mathcal{A}}(a)/|\mathcal{A}|$ to be the uniform distribution over \mathcal{A} .

Step 2 (Global game)

The next step is to merge the local distributions over small neighborhoods into a global distribution over $[K] = \mathcal{A} \cup \mathcal{D}$. A square matrix is **right stochastic** if it has positive entries and its rows sum to one. Such a $d \times d$ matrix describes a homogeneous Markov chain with state-space $[d]$ and row $i \in [d]$ of the matrix defines the distribution over the next-states. We have briefly met homogeneous Markov chains in Section 3.2. The following result is all that we need, the proof of which is to the reader (Exercise 36.8).

LEMMA 36.2 *Let Q_t be the right stochastic matrix with k th row equal to Q_{tk} . Then there exists a unique distribution \tilde{P}_t such that $\tilde{P}_t^\top = \tilde{P}_t^\top Q_t$. Furthermore, this distribution is supported on \mathcal{A} .*

The distribution \tilde{P}_t is called the **stationary distribution** of the Markov chain with kernel Q_t . It is supported on \mathcal{A} because following Q_t never transitions to states outside of \mathcal{A} . The reader may at this point wonder about why were the actions in \mathcal{D} even included in the first place: The answer is that we want \tilde{P}_t to be defined over $[K]$ merely to simplify some expressions that follow. By rewriting the matrix multiplication we see that

$$\tilde{P}_{tk} = \sum_{a \in \mathcal{A}} \tilde{P}_{ta} Q_{tak}, \tag{36.10}$$

which we use repeatedly in the analysis that follows. In particular, this identity plays a key role in relating the regret to a weighted sum of ‘local regrets’.

Step 3 (Redistribution)

Now \tilde{P}_t is rebalanced to a new distribution P_t for which duplicate and degenerate actions in \mathcal{D} are played with sufficient probability. This is done iteratively, starting with $P_t = \tilde{P}_t$. Then for each $d \in \mathcal{D}$ the algorithm finds actions $a, b \in \mathcal{A}$ such that $\ell_d = \alpha\ell_a + (1 - \alpha)\ell_b$ for some $\alpha \in [0, 1]$, which is possible by Lemma 36.1. Then P_t is updated so that some of the probability assigned to actions a and b is transferred to action d . After mass has been assigned to all degenerate actions the algorithm incorporates a small amount of fixed exploration. The complete procedure is given in Algorithm 22. This is done in such a way that the expected loss of playing according to P_t is approximately the same as \tilde{P}_t . The next lemma formalizes the properties of P_t that will be critical in what follows. The proof is left to the reader (Exercise 36.9).

LEMMA 36.3 *Assume $\gamma \in [0, 1/2]$, let $u \in \mathcal{P}_{E-1}$ and let $a, k \in \mathcal{A}$ be arbitrary neighbors. Then $P_t \in \mathcal{P}_{K-1}$ is a probability vector and the following hold:*

- (a) $P_{ta} \geq \tilde{P}_{ta}/4$.
- (b) $\left| \sum_{a=1}^K (P_{ta} - \tilde{P}_{ta}) \langle \ell_a, u \rangle \right| \leq \gamma$.
- (c) $P_{tb} \geq \frac{\tilde{P}_{tk} Q_{tka}}{4K}$ for any non-duplicate $b \in \mathcal{N}_{ka}$.
- (d) $P_{ta} \geq \gamma/K$.
- (e) $P_{td} \geq \frac{\tilde{P}_{tk}}{4K}$ for any $d \in [K]$ such that $\ell_d = \ell_k$.

Step 4 (Acting and estimating)

By the definition of local observability, for each pair of neighboring actions a, b there exists a function $v^{ab} : [K] \times [F] \rightarrow \mathbb{R}$ satisfying Eq. (36.3) and with $v^{ab}(c, f) = 0$ whenever $c \notin \mathcal{N}_{ab}$. Even though a is not a neighbor of itself, for notational convenience we define $v^{aa}(c, f) = 0$ for all c, f . While the policy will work for any admissible choice of v^{ab} , the analysis suggests minimizing

$$V = \max_{a,b} \|v^{ab}\|_\infty$$

with the maximum over all pairs of neighbors.



In Exercise 36.12 you will show that if $|\mathcal{N}_{ab}| = 2$, then v^{ab} can be chosen so that $\|v^{ab}\|_\infty \leq 1 + F$ and that in the worst case this bound is tight. This result no longer holds for larger \mathcal{N}_{ab} as discussed in the exercise.

In round t , the action A_t is chosen at random from P_t . The loss difference estimators are then computed by

$$\hat{Z}_{tka} = \hat{Z}_{tka} - \beta_{tka},$$

where \hat{Z}_{tka} is an unbiased estimator of $y_{ta} - y_{tk}$ and β_{tka} is a bias term:

$$\hat{Z}_{tka} = \frac{\tilde{P}_{tk} v^{ak}(A_t, \Phi_t)}{P_{tA_t}} \quad \text{and} \quad \beta_{tka} = \eta V^2 \sum_{b \in \mathcal{N}_{ak}} \frac{\tilde{P}_{tb}^2}{P_{tb}}. \quad (36.11)$$

The four steps described so far are summarized in Algorithm 22 below.

```

1: Input  $\mathcal{L}, \Phi, \eta, \gamma$ 
2: for  $t \in 1, \dots, n$  do
3:   For  $a, k \in [K]$  let
      
$$Q_{tka} = \mathbb{I}_{\mathcal{A}}(k) \frac{\mathbb{I}_{\mathcal{N}_k \cap \mathcal{A}}(a) \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{ska}\right)}{\sum_{b \in \mathcal{N}_k \cap \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{ska}\right)} + \mathbb{I}_{\mathcal{D}}(k) \frac{\mathbb{I}_{\mathcal{A}}(a)}{|\mathcal{A}|}$$

4:   Find distribution  $\tilde{P}_t$  such that  $\tilde{P}_t^\top = \tilde{P}_t^\top Q_t$ 
5:   Compute  $P_t = (1 - \gamma)\text{REDISTRIBUTE}(\tilde{P}_t) + \frac{\gamma}{K} \mathbf{1}$  and sample  $A_t \sim P_t$ 
6:   Compute loss-difference estimators for each  $k \in \mathcal{A}$  and  $a \in \mathcal{N}_k \cap \mathcal{A}$ .
      
$$\hat{Z}_{tka} = \frac{\tilde{P}_{tk} v^{ak}(A_t, \Phi_t)}{P_{tA_t}}$$

      
$$\beta_{tka} = \eta V^2 \sum_{b \in \mathcal{N}_{ak}} \frac{\tilde{P}_{tb}^2}{P_{tb}} \quad (36.12)$$

      
$$\tilde{Z}_{tka} = \hat{Z}_{tka} - \beta_{tka}$$

7: end for
8: function  $\text{REDISTRIBUTE}(p)$ 
9:    $q \leftarrow p$ 
10:  for  $d \in \mathcal{D}$  do
11:    Find  $a, b$  with  $d \in \mathcal{N}_{ab}$  and  $\alpha \in [0, 1]$  such that  $\ell_d = \alpha \ell_a + (1 - \alpha) \ell_b$ 
12:     $c_a \leftarrow \frac{\alpha q_b}{\alpha q_b + (1 - \alpha) q_a}$  and  $c_b \leftarrow 1 - c_a$  and  $\rho \leftarrow \frac{1}{2K} \min \left\{ \frac{p_a}{q_a c_a}, \frac{p_b}{q_b c_b} \right\}$ 
13:     $q_d \leftarrow \rho c_a q_a + \rho c_b q_b$  and  $q_a \leftarrow (1 - \rho c_a) q_a$  and  $q_b \leftarrow (1 - \rho c_b) q_b$ 
14:  end for
15:  return  $q$ 
16: end function
```

Algorithm 22: NeighborhoodWatch2

The next theorem bounds the regret of NeighborhoodWatch2 with high probability for locally observable games.

THEOREM 36.5 *Let \hat{R}_n be the random regret*

$$\hat{R}_n = \max_{b \in [K]} \sum_{t=1}^n \langle \ell_{A_t} - \ell_b, u_t \rangle.$$

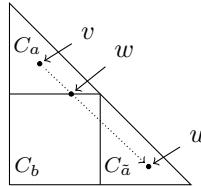


Figure 36.3 The construction used in the proof of Lemma 36.4.

Suppose that Algorithm 22 is run on locally observable $G = (\mathcal{L}, \Phi)$ and

$$\eta = \frac{1}{V} \sqrt{\frac{\log(K/\delta)}{nK}} \quad \text{and} \quad \gamma = VK\eta.$$

Let $0 < \delta < 1$. Then with probability at least $1 - \delta$ the regret is bounded by $\hat{R}_n \leq C_G \sqrt{n \log(e/\delta)}$, where C_G is a constant depending on G , but not n , or δ .

By choosing $\delta = 1/n$ the following corollary is obtained.

COROLLARY 36.1 Suppose that Algorithm 22 is run on locally observable $G = (\mathcal{L}, \Phi)$ with the same choices of η and γ as Theorem 36.5 and $\delta = 1/n$, then there exists a constant C'_G depending on G , but not n such that

$$R_n \leq C'_G \sqrt{n \log(n)}.$$

36.6 Upper bound for easy games

The first step is a simple lemma showing the regret can be localised to the neighborhood of the played action.

LEMMA 36.4 There exists a constant $\varepsilon_G > 0$ depending only on G such that for all pairs of actions $a, \tilde{a} \in \mathcal{A}$ and $u \in C_{\tilde{a}}$ there exists an action $b \in \mathcal{N}_a \cap \mathcal{A}$ such that $\langle \ell_a - \ell_{\tilde{a}}, u \rangle \leq \langle \ell_a - \ell_b, u \rangle / \varepsilon_G$.

Proof Since $u \in C_{\tilde{a}}$, $0 \leq \langle \ell_a - \ell_{\tilde{a}}, u \rangle$. The result is trivial if a, \tilde{a} are neighbors or $\langle \ell_a - \ell_{\tilde{a}}, u \rangle = 0$. From now on assume that $\langle \ell_a - \ell_{\tilde{a}}, u \rangle > 0$ and that a, \tilde{a} are not neighbors. Let v be the centroid of C_a . The idea is to choose $b \in \mathcal{N}_a \cap \mathcal{A}$ as that neighbor of a whose cell is the one that the line segment that connects v and u enters when leaving C_a . To be precise, if w lies in the intersection of the line segment connecting v and u and the boundary of C_a then b is a neighbor of a in \mathcal{A} so that $w \in C_a \cap C_b$. Note that w is well-defined by the Jordan-Brouwer separation theorem (see the notes at the end of the chapter), and b is well-defined because \mathcal{A} is a maximal duplicate-free subset of the Pareto optimal actions. Using

twice that $\langle \ell_a - \ell_b, w \rangle = 0$, we calculate

$$\begin{aligned} \langle \ell_a - \ell_b, u \rangle &= \langle \ell_a - \ell_b, u - w \rangle = \frac{\|u - w\|_2}{\|v - w\|_2} \langle \ell_a - \ell_b, w - v \rangle \\ &= \frac{\|u - w\|_2}{\|v - w\|_2} \langle \ell_b - \ell_a, v \rangle > 0, \end{aligned} \quad (36.13)$$

where the second equality used that $w \neq v$ is a point of the line segment connecting v and u , hence $w - v$ and $u - w$ are parallel and share the same direction and $\|v - w\|_2 > 0$ (see Fig. 36.3), and the last inequality follows because v is the centroid of C_a and a, b are distinct Pareto optimal actions.

Let v_c be the centroid of C_c for any $c \in \mathcal{A}$. Then,

$$\begin{aligned} \frac{\langle \ell_a - \ell_{\tilde{a}}, u \rangle}{\langle \ell_a - \ell_b, u \rangle} &= \frac{\langle \ell_a - \ell_{\tilde{a}}, w + u - w \rangle}{\langle \ell_a - \ell_b, u \rangle} \stackrel{(a)}{\leq} \frac{\langle \ell_a - \ell_b, w \rangle + \langle \ell_a - \ell_{\tilde{a}}, u - w \rangle}{\langle \ell_a - \ell_b, u \rangle} \\ &\stackrel{(b)}{=} \frac{\langle \ell_a - \ell_{\tilde{a}}, u - w \rangle}{\langle \ell_a - \ell_b, u \rangle} \stackrel{(c)}{=} \frac{\|v - w\|_2 \langle \ell_a - \ell_{\tilde{a}}, u - w \rangle}{\|u - w\|_2 \langle \ell_b - \ell_a, v \rangle} \\ &\stackrel{(d)}{\leq} \frac{\|v - w\|_2 \|\ell_a - \ell_{\tilde{a}}\|_2}{\langle \ell_b - \ell_a, v \rangle} \stackrel{(e)}{\leq} \frac{\sqrt{2E}}{\min_{c \in \mathcal{A}} \min_{d \in \mathcal{N}_c} \langle \ell_d - \ell_c, v_c \rangle} = \frac{1}{\varepsilon_G}, \end{aligned}$$

where (a) follows since by (36.13), $\langle \ell_a - \ell_b, u \rangle > 0$ and also because $w \in C_b$ implies that $\langle \ell_a - \ell_{\tilde{a}}, w \rangle \leq \langle \ell_a - \ell_b, w \rangle$, (b) follows since $\langle \ell_a - \ell_b, w \rangle = 0$, which is used in other steps as well. (c) uses (36.13), (d) is by Cauchy-Schwartz and in (e) we bounded $\|w - v\|_2 \leq \sqrt{2}$ and used that $\|\ell_a - \ell_{\tilde{a}}\|_2 \leq \sqrt{E}$ and $\langle \ell_b - \ell_a, v \rangle = \langle \ell_b - \ell_a, v_a \rangle \geq \min_{c \in \mathcal{A}} \min_{d \in \mathcal{N}_c} \langle \ell_d - \ell_c, v_c \rangle > 0$. The final equality serves as the definition of $1/\varepsilon_G$. \square

LEMMA 36.5 *Let \mathcal{H} be the set of functions $\phi : \mathcal{A} \rightarrow \mathcal{A}$ with $\phi(a) \in \mathcal{N}_a$ for all $a \in \mathcal{A}$ and define $a_n^* = \operatorname{argmin}_{a \in [K]} \sum_{t=1}^n \langle \ell_a, u_t \rangle$. Then, for any $(B_t)_{1 \leq t \leq n}$ sequence of actions in \mathcal{A} ,*

$$\sum_{t=1}^n \langle \ell_{B_t} - \ell_{a_n^*}, u_t \rangle \leq \frac{1}{\varepsilon_G} \max_{\phi \in \mathcal{H}} \sum_{t=1}^n \langle \ell_{B_t} - \ell_{\phi(B_t)}, u_t \rangle.$$

Lemma 36.5 With no loss of generality, we can assume that $a_n^* \in \mathcal{A}$ because \mathcal{A} is a maximal duplicate-free subset of Pareto optimal actions. Apply the previous lemma on subsequences of rounds where $B_t = a$ for each $a \in \mathcal{A}$. \square

LEMMA 36.6 *Let $\delta \in (0, 1)$. Then with probability at least $1 - 2\delta$ it holds that*

$$\hat{R}_n \leq \gamma n + \frac{1}{\varepsilon_G} \sum_{k \in \mathcal{A}} \max_{b \in \mathcal{N}_k \cap \mathcal{A}} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a \in \mathcal{A}} Q_{tka} (y_{ta} - y_{tb}) + \sqrt{8n \log(|\mathcal{H}|/\delta)}.$$

Proof For $t \in [n]$, let $B_t \sim \tilde{P}_t$. Define the surrogate regret $\hat{R}'_n = \sum_{t=1}^n \langle \ell_{B_t} - \ell_{a_n^*}, u_t \rangle$. By the definition of A_t and B_t and Lemma 36.3 we have $\mathbb{E}_{t-1}[\langle \ell_{A_t} - \ell_{B_t}, u_t \rangle] \leq \gamma$. Furthermore, $|\langle \ell_a - \ell_b, u_t \rangle| \leq 1$ for all a, b . Therefore, by Hoeffding-Azuma, with probability at least $1 - \delta$,

$$\hat{R}_n \leq \hat{R}'_n + \gamma n + \sqrt{2n \log(1/\delta)}. \quad (36.14)$$

By Lemma 36.5, the surrogate regret is bounded in terms of the local regret:

$$\hat{R}'_n = \sum_{t=1}^n \langle \ell_{B_t} - \ell_{a_n^*}, u_t \rangle \leq \frac{1}{\varepsilon_G} \max_{\phi \in \mathcal{H}} \sum_{t=1}^n \langle \ell_{B_t} - \ell_{\phi(B_t)}, u_t \rangle. \quad (36.15)$$

We prepare to use Hoeffding-Azuma again. Fix $\phi \in \mathcal{H}$ arbitrarily. Then,

$$\begin{aligned} \mathbb{E}_{t-1} [\langle \ell_{B_t} - \ell_{\phi(B_t)}, u_t \rangle] &= \sum_{k \in \mathcal{A}} \tilde{P}_{tk} \sum_{a \in \mathcal{A}} Q_{tka} \langle \ell_a - \ell_{\phi(k)}, u_t \rangle \\ &= \sum_{k \in \mathcal{A}} \tilde{P}_{tk} \sum_{a \in \mathcal{A}} Q_{tka} (y_{ta} - y_{t\phi(k)}), \end{aligned}$$

where we used the fact that $\tilde{P}_{ta} = \sum_k \tilde{P}_{tk} Q_{tka}$. Hoeffding-Azuma's inequality now shows that with probability at least $1 - \delta/|\mathcal{H}|$,

$$\sum_{t=1}^n \langle \ell_{B_t} - \ell_{\phi(B_t)}, u_t \rangle \leq \sum_{k \in \mathcal{A}} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a \in \mathcal{A}} Q_{tka} (y_{ta} - y_{t\phi(k)}) + \sqrt{2n \log(|\mathcal{H}|/\delta)}.$$

The result is completed via a union bound over all $\phi \in \mathcal{H}$ and chaining with Eqs. (36.14) and (36.15), and noting that

$$\begin{aligned} \max_{\phi} \sum_{k \in \mathcal{A}} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a \in \mathcal{A}} Q_{tka} (y_{ta} - y_{t\phi(k)}) &\leq \sum_{k \in \mathcal{A}} \max_{\phi} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a \in \mathcal{A}} Q_{tka} (y_{ta} - y_{t\phi(k)}) \\ &= \sum_{k \in \mathcal{A}} \underbrace{\max_{b \in \mathcal{N}_k \cap \mathcal{A}} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a \in \mathcal{A}} Q_{tka} (y_{ta} - y_{tb})}_{\hat{R}_{nk}}. \end{aligned} \quad (36.16)$$

□

Proof of Theorem 36.5 The proof has two steps: Bounding the local regret \hat{R}_{nk} for each $k \in \mathcal{A}$, and then merging the bounds.

Step 1: Bounding the local regret

For the remainder of this step we fix $k \in \mathcal{A}$ and bound the local regret \hat{R}_{nk} . First, we need to massage the local regret into a form in which we can apply the result of Exercise 12.2 in Chapter 12. Let $Z_{tka} = \tilde{P}_{tk} (y_{ta} - y_{tk})$ and \mathcal{G}_t be the σ -algebra generated by (A_1, \dots, A_t) . Let $\mathcal{G} = (\mathcal{G}_t)_{t=0}^n$ be the associated filtration. A simple rewriting shows that

$$\hat{R}_{nk} = \max_{b \in \mathcal{N}_k \cap \mathcal{A}} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a \in \mathcal{A}} Q_{tka} (y_{ta} - y_{tb}) = \max_{b \in \mathcal{N}_k \cap \mathcal{A}} \sum_{t=1}^n \sum_{a \in \mathcal{A}} Q_{tka} (Z_{tka} - Z_{tka}).$$

In order to apply the result in Exercise 12.2 we need to check the conditions. Since $(P_t)_t$ and $(\tilde{P}_t)_t$ are \mathcal{G} -predictable it follows that $(\beta_t)_t$ and $(Z_t)_t$ are also \mathcal{G} -predictable. Similarly, $(\hat{Z}_t)_t$ is \mathcal{G} -adapted because $(A_t)_t$ and $(\Phi_t)_t$ are \mathcal{G} -adapted. It remains to show that assumptions (a-d) are satisfied. For (a) let $a \in \mathcal{N}_k \cap \mathcal{A}$. By part (d) of Lemma 36.3 we have $P_{tb} \geq \gamma/K$ for all t and $b \in [K]$. Furthermore, $|v^{ak}(A_t, \Phi_t)| \leq V$ so that $\eta |\hat{Z}_{tka}| = |\eta \tilde{P}_{tk} v^{ak}(A_t, \Phi_t)/P_{tA_t}| \leq \eta V K / \gamma = 1$, where

the equality follows from the choice of γ . Assumption (b) is satisfied in a similar way with $\eta\beta_{tka} = \eta^2 V^2 \sum_{b \in \mathcal{N}_{ak}} \tilde{P}_{tk}^2 / P_{tb} \leq \eta^2 K^2 V^2 / \gamma = \eta V \leq 1$, where in the last inequality we used the definition of η and assumed that $n \geq \log(K/\delta)$. To make sure that the regret bound holds even for smaller values of n , we require $C_G \geq K\sqrt{\log(eK)}$ so that when $n < K^2 \log(K/\delta)$, the regret bound is trivial. For assumption (c), we have

$$\begin{aligned}\mathbb{E}_{t-1}[\hat{Z}_{tka}^2] &= \mathbb{E}_{t-1} \left[\left(\frac{\tilde{P}_{tk} v^{ak}(A_t, \Phi_t)}{P_{tA_t}} \right)^2 \right] \leq V^2 \tilde{P}_{tk}^2 \mathbb{E}_{t-1} \left[\frac{\mathbb{I}_{\mathcal{N}_{ak}}(A_t)}{P_{tA_t}^2} \right] \\ &= V^2 \sum_{b \in \mathcal{N}_{ak}} \frac{\tilde{P}_{tk}^2}{P_{tb}} = \frac{\beta_{tka}}{\eta}.\end{aligned}$$

Finally (d) is satisfied by the definition of v^{ak} and the fact that $P_t \in \text{ri}(\mathcal{P}_{K-1})$. The result of Exercise 12.2 shows that with probability at least $1 - (K+1)\delta$,

$$\hat{R}_{nk} \leq \frac{3 \log(1/\delta)}{\eta} + 5 \sum_{t=1}^n \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \beta_{tka} + \eta \sum_{t=1}^n \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \hat{Z}_{tka}^2.$$

Step 2: Aggregating the local regret

Using the result from the previous step in combination with a union bound over $k \in \mathcal{A}$ we have that with probability at least $1 - K(K+1)\delta$,

$$\sum_{k \in \mathcal{A}} \hat{R}_{nk} \leq \frac{3K \log(1/\delta)}{\eta} + 5 \sum_{t=1}^n \sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \beta_{tka} + \eta \sum_{t=1}^n \sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{N}_a \cap \mathcal{A}} Q_{tka} \hat{Z}_{tka}^2. \quad (36.17)$$

For bounding the second term we use the definition of β_{tka} from (36.11) and write

$$\sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \beta_{tka} = \eta V^2 \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \sum_{b \in \mathcal{N}_{ak}} \frac{\tilde{P}_{tk}^2}{P_{tb}} = \eta V^2 \tilde{P}_{tk} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \sum_{b \in \mathcal{N}_{ak}} \frac{\tilde{P}_{tk}}{P_{tb}}.$$

We now split the sum that runs over $b \in \mathcal{N}_{ak}$ into two, separating duplicates of k and the rest:

$$\begin{aligned}\sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \sum_{b \in \mathcal{N}_{ak}} \frac{\tilde{P}_{tk}}{P_{tb}} &= \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \sum_{b: \ell_b = \ell_k} \frac{\tilde{P}_{tk}}{P_{tb}} + \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \sum_{b \in \mathcal{N}_{ak}: \ell_b \neq \ell_k} \frac{\tilde{P}_{tk}}{P_{tb}} \\ &= \sum_{b: \ell_b = \ell_k} \frac{\tilde{P}_{tk}}{P_{tb}} + \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} \sum_{b \in \mathcal{N}_{ak}: \ell_b \neq \ell_k} \frac{Q_{tka} \tilde{P}_{tk}}{P_{tb}} \\ &\leq 4K \left(\sum_{b: \ell_b = \ell_k} 1 + \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} \sum_{b \in \mathcal{N}_{ak}: \ell_b \neq \ell_k} 1 \right) \leq 4K^2,\end{aligned} \quad (36.18)$$

where the first equality used that $\sum_a Q_{tka} = 1$, the second to last inequality follows using parts (c) and (e) of Lemma 36.3, stationarity of \tilde{P}_t , and the last

inequality uses a simple counting argument. Details of the arguments needed to show the last two inequalities are left to reader in Exercise 36.10. Summing over all rounds and $k \in \mathcal{A}$ yields

$$5 \sum_{t=1}^n \sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \beta_{tka} \leq 20\eta n K^2 V^2.$$

For the last term in Eq. (36.17) we use the definition of \hat{Z}_{tka} and Parts (c) and (e) of Lemma 36.3 to show that

$$\begin{aligned} \eta \sum_{t=1}^n \sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \hat{Z}_{tka}^2 &= \eta \sum_{t=1}^n \sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} \frac{Q_{tka} \tilde{P}_{tk}^2 v^{ak}(A_t, \Phi_t)^2}{P_{tA_t}^2} \\ &\leq \eta V^2 \sum_{t=1}^n \frac{1}{P_{tA_t}} \sum_{k \in \mathcal{A}} \tilde{P}_{tk} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} \frac{Q_{tka} \tilde{P}_{tk} \mathbb{I}_{\mathcal{N}_{ak}}(A_t)}{P_{tA_t}} \\ &\leq 4\eta K V^2 \sum_{t=1}^n \frac{1}{P_{tA_t}}, \end{aligned}$$

where the last step follows by splitting the sum over a into two based on whether A_t is a duplicate of k and following an argument similar to the one used to prove (36.18). Now, from Part (d) of Lemma 36.3, $\gamma/K(1/P_{ta}) \leq 1$ for all a , and in particular, holds for $a = A_t$. Furthermore, $\mathbb{E}_{t-1}[1/P_{tA_t}] = K$ and $\mathbb{E}_{t-1}[1/P_{tA_t}^2] = \sum_a 1/P_{ta} \leq K^2/\gamma$. By the result in Exercise 5.17 we get that it holds that with probability at least $1 - \delta$,

$$\sum_{t=1}^n \frac{1}{P_{tA_t}} \leq 2nK + \frac{K \log(1/\delta)}{\gamma}.$$

Another union bound shows that with probability at least $1 - (1 + K(K+1))\delta$,

$$\sum_{k \in \mathcal{A}} \hat{R}_{nk} \leq \frac{3K \log(1/\delta)}{\eta} + 28\eta n V^2 K^2 + 4VK \log(1/\delta).$$

The result follows from the definition of η , Lemma 36.6 and the definition of \hat{R}_{nk} . \square

36.7 Proof of the classification theorem

Almost all the results are now available to prove Theorem 36.1. In Section 36.4 we showed that if G is globally observable and not locally observable, then $R_n^*(G) = \Omega(n^{2/3})$. We also proved that if G is locally observable and has neighbors, then $R_n^*(G) = \Omega(\sqrt{n})$. This last result is complemented by the policy and analysis in Section 36.6 where we showed that for locally observable problems $R_n^*(G) = O(\sqrt{n \log(n)})$. Finally we proved that if G is not globally observable, then $R_n^*(G) = \Omega(n)$. All that remains is to prove that (a) if G has no neighboring

actions, then $R_n^*(G) = 0$ and (b) if G is globally observable, but not locally observable, then $R_n^*(G) = O(n^{2/3})$.

THEOREM 36.6 *If G has no neighboring actions, then $R_n^*(G) = 0$.*

Proof Since G has no neighboring actions, there exists an action a such that $C_a = \mathcal{P}_{E-1}$ and the policy that chooses $A_t = a$ for all rounds suffers no regret. \square

THEOREM 36.7 *If G is globally observable, then $R_n^*(G) = O(n^{2/3})$.*

Proof sketch Let $\mathcal{A} \subseteq [K]$ be the set of Pareto optimal actions and $a_\circ \in \mathcal{A}$. Use the definition of global observability to show that each $a \in \mathcal{A}$ there exists a function $h^a : [K] \times [F] \rightarrow \mathbb{R}$ such that

$$\sum_{b=1}^K h^a(b, \Phi(b, i)) = \ell_{ai} - \ell_{a_\circ i} \quad \text{for all } i \in [E].$$

Then define unbiased loss estimator $\hat{\Delta}_{ta} = h^a(A_t, \Phi_t)/P_{tA_t}$, where

$$P_{ta} = (1 - \gamma) \frac{\mathbb{I}_{\mathcal{A}}(a) \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\Delta}_{sa}\right)}{\sum_{b \in \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\Delta}_{sb}\right)} + \frac{\gamma}{K}.$$

The result is completed by repeating the standard analysis of the exponential weights algorithm (or mirror descent with negentropy potential) and optimizing γ and η . \square

36.8 Notes

- 1 A nonempty set $L \subseteq \mathbb{R}^n$ is a **linear subspace** of \mathbb{R}^n if $\alpha v + \beta w \in L$ for all $\alpha, \beta \in \mathbb{R}$ and $v, w \in L$. If L and M are linear subspaces of \mathbb{R}^n , then $L \oplus M = \{v + w : L \in L, w \in M\}$. The **orthogonal complement** of linear subspace L is $L^\perp = \{v \in \mathbb{R}^n : \langle u, v \rangle = 0 \text{ for all } u \in L\}$. The following properties are easily checked: (i) L^\perp is a linear subspace, (ii) $(L^\perp)^\perp = L$ and (iii) $(L \cap M)^\perp = L^\perp \oplus M^\perp$.
- 2 Let $A \in \mathbb{R}^{m \times n}$ be a matrix and recall that matrices of this form correspond to linear maps from $\mathbb{R}^n \rightarrow \mathbb{R}^m$ where the function $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is given by matrix multiplication, $A(x) = Ax$. The **image** of A is $\text{im}(A) = \{Ax : x \in \mathbb{R}^n\}$ and the **kernel** is $\ker(A) = \{x \in \mathbb{R}^n : Ax = 0\}$. Notice that $\text{im}(A) \subseteq \mathbb{R}^m$ and $\ker(A) \subseteq \mathbb{R}^n$. One can easily check that $\text{im}(A)$ and $\ker(A^\top)$ are linear subspaces and an elementary theorem in linear algebra says that $\text{im}(A) \oplus \ker(A^\top) = \mathbb{R}^m$ for any matrix $A \in \mathbb{R}^{m \times n}$. Finally, if $u \in \text{im}(A)$ and $v \in \ker(A^\top)$, then $\langle u, v \rangle = 0$. There are probably hundreds of introductory texts on linear algebra. A short and intuitive exposition is by Axler [1997].

3 Given a set $A \subseteq \mathbb{R}^d$ the **affine hull** is the set

$$\text{aff}(A) = \left\{ \sum_{i=1}^k \alpha_i x_k : k > 0, \alpha \in \mathbb{R}^k, x_i \in A \text{ for all } i \in [k] \text{ and } \sum_{i=1}^k \alpha_i = 1 \right\}.$$

Its dimension is the smallest m such that there exist vectors $v_1, \dots, v_m \in \mathbb{R}^d$ such that $\text{aff}(A) = x_0 + \text{span}(v_1, \dots, v_m)$ for any $x_0 \in A$.

4 We introduced the stochastic variant of partial monitoring to prove our lower bounds. Of course our upper bounds also apply to this setting, which means the classification theorem also holds in the stochastic case. The interesting question is to understand the problem-dependent regret, which for partial monitoring problem $G = (\mathcal{L}, \Phi)$ is

$$R_n(\pi, u, G) = \max_{a \in [K]} \mathbb{E} \left[\sum_{t=1}^n \langle \ell_{A_t} - \ell_a, U_t \rangle \right],$$

where U, U_1, \dots, U_n is a sequence of independent and identically distributed random vectors with $U_t \in \{e_1, \dots, e_E\}$ and $\mathbb{E}[U] = u \in \mathcal{P}_{E-1}$. Provided G is not hopeless one can derive an algorithm for which the regret is logarithmic, and like in bandits there is a sense of asymptotic optimality. The open research question is to understand the in-between regime where the horizon is not yet large enough that the asymptotically optimal logarithmic regret guarantees become meaningful, but not so small that minimax is acceptable.

5 In the proof of Lemma 36.4 we used the overpowered Jordan-Brouwer separation theorem to guarantee that the line segment that connects u with the centroid v of C_a has a nonempty intersection with the boundary of C_a . Here, u was a point that lied outside of C_a . The Jordan-Brouwer separation theorem generalizes the Jordan curve theorem, which states that every simple closed planar curve separates the plane into a bounded interior and an unbounded exterior region so that the boundary of both regions is the said planar curve. The Jordan-Brouwer theorem states that the same holds in higher dimensions where the closed planar curve becomes a topological sphere, which is the image of the unit sphere of \mathbb{R}^d under some continuous injective map from the sphere into \mathbb{R}^d . To use the theorem we view the simplex \mathcal{P}_{E-1} as a subset of \mathbb{R}^{E-1} by dropping the last entry in the coordinate representation of the points of \mathcal{P}_{E-1} . Then the boundary of C_a can be seen as a topological sphere in \mathbb{R}^{E-1} and v belongs to the interior, while u belongs to the exterior region created by the boundary of C_a . The line segment connecting u and v will pass through the boundary of both regions, which happens to be the boundary of C_a , showing that the intersection of the line segment and the boundary of C_a is nonempty. Note that the argument does not show that the intersection has a single point and we did not need this either. Nevertheless, it is not hard to see that this is also true. The standard proof of the Jordan-Brouwer is an application of algebraic topology [Hatcher, 2002, §2.B].

6 Partial monitoring has many potential applications. We already mentioned

dynamic pricing and spam filtering. In the latter case acquiring the true label comes at a price, which is a typical component of hard partial monitoring problems. In general there are many setups where the learner can pay extra for high quality information. For example, in medical diagnosis the doctor can request additional tests before recommending a treatment plan, but these cost time and money. Yet another potential application is quality testing in factory production where the quality control team can choose which items to test (at great cost).

- 7 There are many possible extensions to the partial monitoring framework. We have only discussed problems where the number of actions/feedbacks/outcomes are potentially infinite, but nothing prevents studying a more general setting. Suppose the learner chooses a sequence of real-valued outcomes i_1, \dots, i_n with $i_t \in [0, 1]$. In each round the learner chooses $A_t \in [K]$ and observes $\Phi_{A_t}(i_t)$ where $\Phi_a : [0, 1] \rightarrow \Sigma$ is a known feedback function. The loss is determined by a collection of known functions $\mathcal{L}_a : [0, 1] \rightarrow [0, 1]$. We do not know of any systematic study of this setting. The reader can no doubt imagine generalizing this idea to infinite action sets or introducing a linear structure for the loss.
- 8 A pair of Pareto-optimal actions (a, b) are called **weak neighbors** if $C_a \cap C_b \neq \emptyset$ and **pairwise observable** if there exists a function v satisfying Eq. (36.3) and with $v(c, f) = 0$ whenever $c \notin \{a, b\}$. A partial monitoring problem is called a **point-locally observable game** if all weak neighbours are pairwise observable. All point-locally observable games are locally observable, but the converse is not true. [Bartók \[2013\]](#) designed a policy for this type of game for which

$$R_n \leq \frac{1}{\varepsilon_G} \sqrt{K_{\text{loc}} n \log(n)},$$

where $\varepsilon_G > 0$ is a game-dependent constant and K_{loc} is the size of the largest $A \subseteq [K]$ of Pareto optimal actions such that $\cap_{a \in A} C_a \neq \emptyset$. Using a different policy, [Lattimore and Szepesvári \[2018\]](#) have shown that as the horizon grows the game-dependence diminishes so that

$$\lim_{n \rightarrow \infty} \frac{R_n}{\sqrt{n}} \leq 8(2 + F) \sqrt{2K_{\text{loc}} \log(K)}.$$

- 9 Linear regret is unavoidable in hopeless games, but that does not mean there is nothing to play for. [Rustichini](#) considered a version of the regret that captures the performance of policies in this hard setting. Given $p \in \mathcal{P}_{E-1}$ define set $\mathcal{I}(p) \subseteq \mathcal{P}_{E-1}$ by

$$\mathcal{I}(p) = \left\{ q \in \mathcal{P}_{E-1} : \sum_{i=1}^E (p_i - q_i) \mathbb{I}\{\Phi_{ai} = f\} \text{ for all } a \in [K] \text{ and } f \in [F] \right\}.$$

This is the set of distributions over the outcomes that are indistinguishable

from p by the learner using any actions. Then define

$$f(p) = \max_{q \in \mathcal{I}(p)} \min_{a \in [K]} \sum_{i=1}^E q_i \mathcal{L}_{ai}.$$

[Rustichini \[1999\]](#) proved there exist policies such that

$$\lim_{n \rightarrow \infty} \max_{i_{1:n}} \mathbb{E} \left[\frac{1}{n} \sum_{t=1}^n \mathcal{L}_{A_t i_t} - f(\bar{u}_n) \right] = 0,$$

where $\bar{u}_n = \frac{1}{n} \sum_{t=1}^n e_{i_t} \in \mathcal{P}_{E-1}$ is the average outcome chosen by the adversary. Intuitively this means the learner does not compete with the best action in hindsight with respect to the actual outcomes. Instead, the learner competes with the best action in hindsight with respect to an outcome sequence that is indistinguishable from the actual outcome sequence. [Rustichini](#) did not prove rates on the convergence of the limit. This has been remedied recently and we give some references in the bibliographic remarks.

- 10 Finally, we want to emphasize that partial monitoring is still quite poorly understood. We do not know how the regret should depend on E , F , K or the structure of G . Lower bounds that depend on these quantities are also missing and the lower bounds proven in Section 36.4 are surely very conservative. We hope this chapter inspires more activity in this area. The setting described in the previous note is even more wide open, with even the dependence on n still not completely nailed down.

36.9 Bibliographical remarks

The first work on partial monitoring is by [Rustichini \[1999\]](#), who focussed on the finding Hannan consistent policies in the adversarial setting. [Rustichini](#) shows how to reduce the problem to Blackwell approachability (see [Cesa-Bianchi and Lugosi \[2006\]](#)) and uses this to deduce the existence of a Hannan consistent strategy. [Rustichini](#) also used a slightly different notion of regret, which eliminates the hopeless games. The first nonasymptotic result in the setting of this chapter is due to [Piccolboni and Schindelhauer \[2001\]](#) where a policy with regret $O(n^{3/4})$ is given for problems that are not hopeless. [Cesa-Bianchi et al. \[2006\]](#) reduced the dependence to $O(n^{2/3})$ and proved a wide range of other results for specific classes of problems. The classification theorem when $E = 2$ is due to [Bartók et al. \[2010\]](#) (extended version: [Antos et al. \[2013\]](#)). The classification of general partial monitoring games is by [Bartók et al. \[2014\]](#). The neighborhood watch policy is due to [Foster and Rakhlin \[2012\]](#). The policy presented here is a simplification of that algorithm [[Lattimore and Szepesvári, 2018](#)]. The policies mentioned in Note 8 are due to [Bartók \[2013\]](#) and [Lattimore and Szepesvári \[2018\]](#). We warn the reader that neighbors are defined differently by [Foster and Rakhlin \[2012\]](#) and [Bartók \[2013\]](#), which can lead to confusion. Additionally, although both papers

are largely correct, in both cases the core proofs contain errors that cannot be resolved without changing the policies [Lattimore and Szepesvári, 2018]. There is also a growing literature on the stochastic setting where it is common to study both minimax and asymptotic bounds. In the latter case one can obtain asymptotically optimal logarithmic regret for games that are not hopeless. We refer the reader to papers by Bartók et al. [2012], Vanchinathan et al. [2014], Komiyama et al. [2015b] as a good starting place. As we mentioned, partial monitoring can model problems that lie between bandits and full information. There are now several papers on this topic, but in more restricted settings and consequentially with more practical algorithms and bounds. One such model is when the learner is playing actions corresponding to vertices on a graph and observes the losses associated with the chosen vertex and its neighbours [Mannor and Shamir, 2011, Alon et al., 2013]. A related result is in the finite-armed Gaussian setting where the learner selects an action $A_t \in [K]$ and observes a Gaussian sample from each arm, but with variances depending on the chosen action. Like partial monitoring this problem exhibits many challenges and is not yet well understood [Wu et al., 2015]. We mentioned in Note 9 that for hopeless games the definition of the regret can be refined. A number of authors have studied this setting with sublinear regret guarantees. As usual, the price of generality is that the bounds are correspondingly a bit worse [Perchet, 2011, Mannor and Shimkin, 2003, Mannor et al., 2014].

36.10 Exercises

36.1 Let $\mathcal{X} \subseteq \mathcal{Y} \subseteq \mathbb{R}^d$ and $\dim(\mathcal{X}) = \dim(\mathcal{Y})$. Prove that $\text{aff}(\mathcal{X}) = \text{aff}(\mathcal{Y})$.

36.2 Calculate the neighborhood structure, cell decomposition and action classification for each of the examples in this chapter.

36.3 Apples arrive sequentially from the farm to a processing facility. Most apples are fine, but occasionally there is a rotten one. The only way to figure out whether an apple is good or rotten is to taste it. For some reason customers do not like bite-marks in the apples they buy, which means that tested apples cannot be sold. Good apples yield a unit reward when sold, while the sale of a bad apple costs the company $c > 0$.

- (a) Formulate this problem as a partial monitoring problem: Determine \mathcal{L} and Φ .
- (b) What is the minimax regret in this problem?
- (c) What do you think about this problem? Will actual farmers be excited about your analysis?

36.4 Let $G = (\mathcal{L}, \Phi)$ be a partial monitoring game with $K = 2$ actions. Prove that G is either trivial, hopeless or easy.

36.5 Complete the last step in the proof of Theorem 36.2.

36.6 Prove Theorem 36.4.

36.7 Prove Theorem 36.3.

36.8 In this exercise you will prove the existence of a stationary distribution. Let $P \in [0, 1]^{d \times d}$ be right stochastic and $A_n = \frac{1}{n} \sum_{t=0}^{n-1} P^t$. Show that:

- (a) A_n is right stochastic.
- (b) $A_n + \frac{1}{n}(P^n - I) = A_n P = P A_n$.
- (c) $P^* = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} P^t$ exists.
- (d) $P^* P = P P^* = P^* P^* = P^*$.
- (e) There exists a stationary distribution.
- (f) Prove Lemma 36.2.



For Parts (c) and (d) you will likely find it useful that the space of right stochastic matrices is compact. Then show that all cluster points of (A_n) are the same.

36.9 Prove Lemma 36.3.

36.10 Prove the last two inequalities shown in Eq. (36.18). In particular, let $k \in \mathcal{A}$ and show that:

- (a) For any $b \in [K]$ such that $\ell_b = \ell_k$, $\tilde{P}_{tk}/P_{tb} \leq 4K$;
- (b) For any $a \in \mathcal{N}_k \cap \mathcal{A}$, $b \in \mathcal{N}_{ak}$ such that $\ell_b \neq \ell_k$, $Q_{tka}\tilde{P}_{tk}/P_{tb} \leq 4K$;
- (c) The sets $S = \{b \in [K] : \ell_b = \ell_k\}$ and the sets $S_a = \{b \in [K] : b \in \mathcal{N}_{ak}, \ell_b \neq \ell_k\}$ where $a \in \mathcal{N}_k \cap \mathcal{A}$ are all disjoint. Hence,

$$\sum_{b: \ell_b = \ell_k} 1 + \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} \sum_{b \in \mathcal{N}_{ak}: \ell_b \neq \ell_k} 1 \leq K.$$

- (d) Put things together and show that the bound of Eq. (36.18) indeed holds.

36.11 Complete the details to prove Theorem 36.7.

36.12 Suppose that a and b are globally observable and let $v : [K] \times [F] \rightarrow \mathbb{R}$ be a function satisfying Eq. (36.3).

- (a) Show that if a, b are pairwise observable, then v can be chosen so that $\|v\|_\infty \leq 1 + F$.
- (b) Next let $F = 2$ and construct a game and pair of actions a, b (not pairwise observable) such that for all v satisfying Eq. (36.3), $\|v\|_\infty \geq c^K$ for constant $c > 1$.

36.13 Consider $G = (\mathcal{L}, \Phi)$ given by

$$\mathcal{L} = \begin{pmatrix} 1 & 0 & 1 & 0 & \cdots & 1 & 0 \\ 0 & 1 & 0 & 1 & \cdots & 0 & 1 \end{pmatrix} \quad \text{and}$$

$$\Phi = \begin{pmatrix} 1 & 2 & 2 & 3 & 3 & 4 & \cdots & F-1 & F-1 & F \\ 1 & 1 & 2 & 2 & 3 & 3 & \cdots & F-2 & F-1 & F-1 \end{pmatrix}.$$

- (a) Show this game is locally observable.
- (b) Prove there exists a universal constant $c > 0$ such that $R_n^*(G) \geq c(F-1)\sqrt{n}$.



The source for previous exercise is the paper by the authors [[Lattimore and Szepesvári, 2018](#)].

36.14 Complete the necessary modification of Lemma 15.1 to show that Eq. (36.7) is true.

36.15 Write a program that accepts as input matrices \mathcal{L} and Φ and outputs the classification of the game.

36.16 In this experiment we test NeighborhoodWatch2 empirically on the spam game with stochastic adversary.

- (a) Implement NeighborhoodWatch2.
- (b) Apply your algorithm to the spam game for a variety of choices of c and stochastic adversary. Try to stress your algorithm as much as possible (for each c choose the most challenging u).
- (c) Plot your results from the previous part. Tell an interesting story.

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

37 Markov Decision Processes

Bandit environments are a sensible model for many simple problems, but they do not model more complex environments where actions have long-term consequences. A brewing company needs to plan ahead when ordering ingredients and the decisions made today affect their position to brew the right amount of beer in the future. A student learning mathematics benefits not only from the immediate reward of learning an interesting topic, but also from their improved job prospects.

A **Markov decision process** is a simple way to incorporate long-term planning into the bandit framework. Like in bandits, the learner chooses actions and receives rewards. But they also observe a **state** and the rewards for different actions depend on the state. Furthermore, the actions chosen affect which state will be observed next.

37.1 Problem setup

A Markov decision process (MDP) is a tuple $M = (\mathcal{S}, \mathcal{A}, P, r)$ that describes the environment. The first two items \mathcal{S} and \mathcal{A} are sets called the **state space** and **action space** respectively and $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$ are their sizes, which may be infinite. An MDP is finite if $S, A < \infty$. The quantity $P = (P_a : s \in \mathcal{S}, a \in \mathcal{A})$ is called the **transition function** with $P_a : \mathcal{S} \times \mathcal{S} \rightarrow [0, 1]$ so that $P_a(s, s')$ is the probability that the learner transitions from state s to s' when taking action a . The last element in the tuple is the reward $r = (r_a : a \in \mathcal{A})$, which is a collection of **reward functions** with $r_a : \mathcal{S} \rightarrow [0, 1]$. When the learner takes action a in state s it receives a deterministic reward of $r_a(s)$. Depending on the situation, the transition and reward functions are often represented as vectors or matrices. When the state space is finite we may assume without loss of generality that $\mathcal{S} = [S]$. We write $P_a(s) \in [0, 1]^S$ as the probability vector with s 'th coordinate given by $P_a(s, s')$. In the same way we let $P_a \in [0, 1]^{S \times S}$ be the right stochastic matrix with $(P_a)_{s,s'} = P_a(s, s')$. Finally, we view r_a as a vector in $[0, 1]^S$ in the natural way.



While we use the same action-set in different states $s, s' \in \mathcal{S}$, this does not mean that $P_a(s)$ or $r_a(s)$ has any relationship to $P_a(s')$ or $r_a(s')$. By learning

about P_a at s the learner does not gain information about P_a at $s' \neq s$. In this sense the notation is a bit misleading and perhaps it would be better to use an entirely different set of actions for each state. This could be done with no changes to any of the results we present. And while we are at it, of course one could also allow the number of actions to vary over the state space. The only justification for assuming that the same set of actions is available in all states is that it simplifies the presentation.

The interaction protocol is very similar to bandits. Before the game starts the initial state S_1 is sampled from a distribution $\mu \in \mathcal{P}(\mathcal{S})$. In each round t the learner observes the state $S_t \in \mathcal{S}$, chooses an action $A_t \in \mathcal{A}$ and receives reward $r_{A_t}(S_t)$. The environment then samples S_{t+1} from the probability vector $P_{A_t}(S_t)$ and then the next round begins (Fig. 37.1).

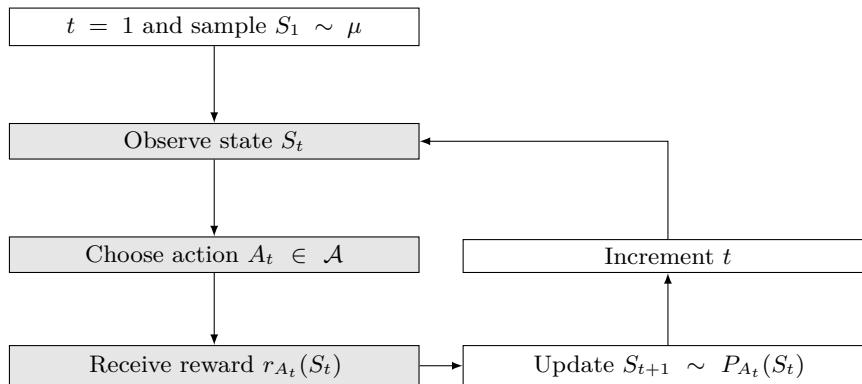


Figure 37.1 Interaction protocol for Markov decision processes

Histories and policies

The **history** $H_t = (S_1, A_1, \dots, S_{t-1}, A_{t-1}, S_t)$ in round t contains the information available before the action for the round is to be chosen. Note that state S_t is included in H_t . The rewards are omitted because they are determined by the state/action pairs and the learner can just recompute them if needed. A policy is a (possibly randomized) map from the set of possible histories to actions. Simple policies include **memoryless policies**, which choose actions based on only the current state, possibly in a randomized manner. The set of such policies is denoted by Π_M and its elements are identified maps $\pi : \mathcal{S} \times \mathcal{A} \times [0, 1]$ with $\sum_{a \in \mathcal{A}} \pi(s, a) = 1$ for any $s \in \mathcal{S}$ so that $\pi(s, a)$ is interpreted as the probability that policy π takes action a in state s .

A memoryless policy that does not randomize is called a **memoryless deterministic policy**. To reduce clutter such policies are written as $\mathcal{S} \rightarrow \mathcal{A}$ maps and the set of all such policies is denoted by Π_{DM} . A policy is called a

Markov policy if the actions are randomized and depend only on the round and the previous state. These policies are represented by fixed sequences of memoryless deterministic policies. Under a Markov policy the sequence of states (S_1, S_2, \dots) evolve as a Markov chain (see Section 3.2). If the Markov policy is memoryless, the chain is homogeneous.

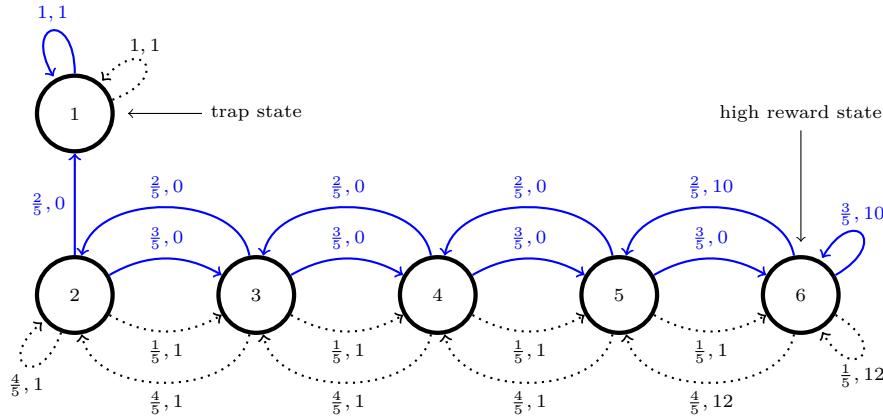


Figure 37.2 A Markov decision process with six states and two actions represented by solid and dashed arrows respectively. The numbers next to each arrow represent the probability of transition and reward for the action respectively. For example, taking the solid action in state three results in a reward of zero and the probability of moving to state four is 3/5 and the probability of moving to state three is 2/5.

Probability spaces

It will be convenient to allow infinitely long interactions between the learner and environment. In line with Fig. 37.1, when the agent or learner follows a policy π in MDP $M = (\mathcal{S}, \mathcal{A}, P, r)$ such a never ending interaction should give rise to a random process $(S_1, A_1, S_2, A_2, \dots)$ so that for any $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$ and $t \geq 1$,

- (a) $\mathbb{P}(S_1 = s) = \mu(s)$;
- (b) $\mathbb{P}(S_{t+1} = s' | H_t, A_t) = P_{A_t}(S_t, s')$;
- (c) $\mathbb{P}(A_t = a | H_t) = \pi(H_t, a)$,

where $\mu \in \mathcal{P}(\mathcal{S})$ is the initial state distribution and $\pi(H_t, a)$ stands for the probability of the agent selecting action a in the t th round of interaction when the history is $H_t = (S_1, A_1, \dots, S_{t-1}, A_{t-1}, S_t)$. At this point, meticulous readers may wonder about whether it is even true that there exist some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sequence of random variables $(S_1, A_1, S_2, A_2, \dots)$ exist at all that make (a)–(c) hold regardless the choice of M , π and μ . This may look like nitpicking, but if this was not guaranteed, all that comes later in this chapter would be vacuous. Our readers should find it pleasing that the Ionescu Tulcea theorem (Theorem 3.3) furnishes us with a positive answer (Exercise 37.1).

Item (b) above is known as the **Markov property**. Of course the measure \mathbb{P} depends on both the policy and Markov decision process. For most of the chapter these quantities will be fixed and the dependence is omitted from the notation. In the few places where disambiguation is necessary we provide additional notation. The initial state distribution μ usually does not play a big role and we allow ourselves to write $\mathbb{P}(\cdot | S_1 = s)$, which just means replacing μ with an alternative initial state distribution that is a Dirac at s .

Traps and the diameter of a Markov decision process

A significant complication in MDPs is the potential for traps. A trap is a subset of the state space that there is no escape from. For example, the MDP in Fig. 37.2 has a trap state. If being in the trap has a suboptimal yield in terms of the reward, the learner should avoid the trap, but since the learner can only discover that an action leads to a trap by trying that action and since, by definition, there are no second chances (the environment-agent interaction is continuous and is uninterrupted, with no option to somehow reset the environment), the problem of learning while competing with a fully informed agent is hopeless (Exercise 37.18).

To avoid this complication we restrict our attention to MDPs with no traps. Formally, we assume that for any pair of states $s, s' \in \mathcal{S}$ there exists a policy such that when starting from s there is a positive probability of reaching s' some time in the future while following the policy. MDPs with this property are called **strongly connected** or **communicating**. One can also define a real-valued measure of the connectedness of an MDP called the **diameter**. MDPs with smaller diameter are usually easier to learn because a policy can recover from mistakes more quickly.

DEFINITION 37.1 Define stopping time $\tau_s = \min\{t \geq 1 : S_t = s\}$. The **diameter** of M is

$$D(M) = \max_{s \neq s'} \min_{\pi: \mathcal{S} \rightarrow \mathcal{A}} \mathbb{E}^{\pi} [\tau_{s'} | S_1 = s] ,$$

where the expectation is taken with respect to the measure on sequences of state/action/reward tuples induced by the interaction with Markov decision process M and policy π .

A number of observations are in order about this definition. First, the order of the maximum and minimum means that for any pair of states a different policy may be used. Second, travel times are always minimized by deterministic memoryless policies so the restriction to these policies in the minimum is inessential (Exercise 37.3). Finally, the definition only considers distinct states. We also note that when the number of states is finite it holds that $D(M) < \infty$ if and only if M is strongly connected (Exercise 37.4).

37.2 Optimal policies and the Bellman optimality equation

We now define the notion of an optimal policy and outline the proof that there exists a deterministic memoryless optimal policy. Throughout we fix a strongly connected Markov decision process M . The **gain** of a policy π is the long-term average reward expected from using that policy when starting in state s :

$$\rho_s^\pi = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n \mathbb{E}^\pi[r_{A_t}(S_t) \mid S_1 = s].$$

In general the limit need not exist, in which case the following quantity can be meaningful:

$$\bar{\rho}_s^\pi = \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n \mathbb{E}^\pi[r_{A_t}(S_t) \mid S_1 = s].$$

Whenever ρ_s^π exists we have $\rho_s^\pi = \bar{\rho}_s^\pi$. The **optimal gain** is a real value

$$\rho^* = \max_{s \in \mathcal{S}} \sup_{\pi} \bar{\rho}_s^\pi,$$

where the supremum is taken over all policies. A π policy is an **optimal policy** if $\rho^\pi = \rho^* \mathbf{1}$. The existence of an optimal policy is far from trivial and we will spend the next little while sketching the proof. You might be wondering why the optimal value does not depend on the initial state. The reason is because we have assumed the Markov decision process is strongly connected so that the learner can travel from one state to any other in a number of rounds that is finite in expectation. The loss suffered during this transition is not captured by the asymptotic definition of the gain.

Before continuing we need some new notation. For memoryless policy π define

$$P_\pi(s, s') = \sum_a \pi(s, a) P_a(s, s') \quad \text{and} \quad r_\pi(s) = \sum_a \pi(s, a) r_a(s). \quad (37.1)$$

We view P_π as an $S \times S$ **transition matrix** and r_π as a vector in \mathbb{R}^S . With this notation P_π is the transition matrix of the homogeneous Markov chain S_1, S_2, \dots when $A_t \sim \pi(S_t, \cdot)$. The gain of memoryless policy π satisfies

$$\rho^\pi = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} P_\pi^t r_\pi = P_\pi^* r_\pi, \quad (37.2)$$

where $P_\pi^* = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} P_\pi^t$ is called the **stationary transition matrix**, the existence of which you will prove in Exercise 37.9. For each $k \in \mathbb{N}$ define

$$v_\pi^{(k)} = \sum_{t=0}^k P_\pi^t (r_\pi - \rho^\pi).$$

For $s \in \mathcal{S}$, $v_\pi^{(k)}(s)$ gives the total expected excess reward collected by π when the process starts at state s . The **(differential) value function** of a policy is a

function $v_\pi : \mathcal{S} \rightarrow \mathbb{R}$ defined as the **Cesàro sum** of the sequence $\{P_\pi^t(r_\pi - \rho^\pi)\}_{t \geq 0}$,

$$v_\pi = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} v_\pi^{(k)} = ((I - P_\pi + P_\pi^*)^{-1} - P_\pi^*)r_\pi. \quad (37.3)$$

Note, the second equality is nontrivial (Exercise 37.9). The definition implies that $v_\pi(s) - v_\pi(s')$ is the ‘average’ long-term advantage of starting in state s relative to starting to state s' when following policy π . Note these quantities are only defined for memoryless policies where they are also guaranteed to exist (Exercise 37.9). Combining Eq. (37.2) and Eq. (37.3) shows that for any memoryless policy π ,

$$\rho^\pi + v_\pi = r_\pi + P_\pi v_\pi.$$

A **value function** is a function $v : \mathcal{S} \rightarrow \mathbb{R}$ and its **span** is given by

$$\text{span}(v) = \max_{s \in \mathcal{S}} v(s) - \min_{s \in \mathcal{S}} v(s).$$

As with other quantities, value functions are associated with vectors in \mathbb{R}^S . A **greedy policy** with respect to value function v is a deterministic memoryless policy π_v given by

$$\pi_v(s) = \operatorname{argmax}_{a \in \mathcal{A}} r_a(s) + \langle P_a(s), v \rangle.$$

There may be many policies that are greedy with respect to some value function v due to ties in the maximum. Usually the ties do not matter, but for consistency and for the sake of simplifying matters, we assume that ties are broken in a systematic fashion. In particular, this makes π_v well-defined for any value function.

One way to find the optimal policy is as the greedy policy with respect to a value function that satisfies the **Bellman optimality equation**, which is

$$\rho + v(s) = \max_{a \in \mathcal{A}} (r_a(s) + \langle P_a(s), v \rangle) \quad \text{for all } s \in \mathcal{S}. \quad (37.4)$$

This is a system of S nonlinear equations with unknowns $\rho \in \mathbb{R}$ and $v \in \mathbb{R}^S$. The reader will notice that if $v : \mathcal{S} \rightarrow \mathbb{R}$ is a solution to Eq. (37.4), then so is $v + c\mathbf{1}$ for any constant $c \in \mathbb{R}$ and hence the Bellman optimality equation lacks unique solutions. Furthermore, it is *not* true that the optimal value function is unique up to translation, even when M is strongly connected (Exercise 37.13). The v -part of a solution pair (ρ, v) of Eq. (37.4) is called an **optimal (differential) value function**.

THEOREM 37.1 *The following hold:*

- (a) *There exists a pair (ρ, v) that satisfies the Bellman optimality equation.*
- (b) *If (ρ, v) satisfies the Bellman optimality equation, then π_v is optimal.*
- (c) *There exists a deterministic memoryless optimal policy.*

Proof sketch The proof of Part (a) is too long to include here, but we guide you

through it in Exercise 37.12. For Part (b) let (ρ, v) satisfy the Bellman equation and π^* be the greedy policy with respect to v . Then

$$\rho^{\pi^*} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} P_{\pi^*}^t r_{\pi^*} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} P_{\pi^*}^t (\rho \mathbf{1} + v - P_{\pi^*} v) = \rho.$$

Next let $\pi = (\pi_1, \pi_2, \dots)$ be an arbitrary Markov policy and $P_\pi^t = \prod_{s=1}^t P_{\pi_s}$. Then using the fact that π^* is the greedy policy with respect to v leads to

$$\begin{aligned} P_\pi^{t-1} r_{\pi_t} &= P_\pi^{t-1} (r_{\pi_t} + P_{\pi_t} v - P_{\pi_t} v) \\ &\leq P_\pi^{t-1} (r_{\pi^*} + P_{\pi^*} v - P_{\pi_t} v) \\ &= P_\pi^{t-1} (\rho \mathbf{1} + v - P_{\pi_t} v) \\ &= \rho \mathbf{1} + P_\pi^{t-1} v - P_\pi^t v. \end{aligned}$$

Summing over t shows that

$$\bar{\rho}^\pi = \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} P_\pi^{t-1} r_{\pi_t} \leq \rho \mathbf{1}.$$

Hence $\rho \geq \bar{\rho}^\pi$ for all Markov policies π . The result is completed using the result of Exercise 37.2, where you will prove that for any policy π there exists a Markov policy with the same expected rewards. Part (c) follows immediately from the first two parts. \square

The theorem shows that there exist solutions to the Bellman optimality equation and that the greedy policy with respect to the resulting value function is an optimal policy. We need one more result about solutions to the Bellman optimality equation, the proof of which you will provide in Exercise 37.14.

LEMMA 37.1 *Suppose that (ρ^*, v) satisfies the Bellman optimality equation. Then $\text{span}(v) \leq D(M)$.*



The operator $T : \mathbb{R}^S \rightarrow \mathbb{R}^S$ defined by $(Tv)(s) = \max_{a \in \mathcal{A}} r_a(s) + \langle P_a(s), v \rangle$ is called the **Bellman operator**. If (ρ^*, v) is a solution to the Bellman optimality equation, then $Tv = \rho^* \mathbf{1} + v$. Furthermore, $v_n^* = T^n \mathbf{0}$ is a vector with $v_n^*(s)$ the maximum achievable expected cumulative reward over n rounds when starting in state s . Value iteration is a procedure for finding the optimal policy that works by starting with an arbitrary value function v_0 and incrementally updating with $v_{k+1} = Tv_k$. Under certain conditions the greedy policy with respect to v_k converges to an optimal policy as k tends to infinity. For more on this see the notes.

We have not said how to solve the Bellman optimality equation. When computation is important this becomes surprisingly subtle. Readers who are more interested in the learning aspect of the problem can skip the details, which are provided in the next section.

37.3 Finding an optimal policy (\dagger)

There are many ways to find an optimal policy, including value iteration, policy iteration and enumeration. These ideas are briefly discussed in the notes. Here we describe an approach based on linear programming. As in the previous section we fix a strongly connected finite Markov decision process. Consider the following linear optimization problem.

$$\begin{aligned} & \underset{\rho \in \mathbb{R}, v \in \mathbb{R}^S}{\text{minimize}} \quad \rho \\ & \text{subject to } \rho + v(s) \geq r_a(s) + \langle P_a(s), v \rangle \text{ for all } s, a. \end{aligned} \tag{37.5}$$

THEOREM 37.2 *The optimization problem in Eq. (37.5) is feasible and if (ρ, v) is a solution, then $\rho = \rho^*$ is the optimal gain.*

Proof By Theorem 37.1 there exists a pair (ρ^*, v^*) satisfying the Bellman optimality equation, which means that for all state/action pairs (s, a) ,

$$\rho^* + v^*(s) = \max_{a \in \mathcal{A}} r_a(s) + \langle P_a(s), v^* \rangle.$$

Hence this pair satisfy the constraints in Eq. (37.5) and witness feasibility. Let (ρ, v) be a solution of Eq. (37.5). Since (ρ^*, v^*) satisfy the constraints, $\rho \leq \rho^*$ is immediate. It remains to prove that $\rho \geq \rho^*$. Let $\pi = \pi_v$ be the greedy policy with respect to v . Then

$$P_{\pi^*}^t r_{\pi^*} \leq P_{\pi^*}^t (r_\pi + P_\pi v - P_{\pi^*} v) \leq P_{\pi^*}^t (\rho \mathbf{1} + v - P_{\pi^*} v) = \rho \mathbf{1} + P_{\pi^*}^t v - P_{\pi^*}^{t+1} v.$$

Summing over t shows that $\rho^* \mathbf{1} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} P_{\pi^*}^t r_{\pi^*} \leq \rho \mathbf{1}$, which shows that $\rho \geq \rho^*$ and completes the proof. \square

We have not claimed that solutions to this linear program satisfy the Bellman optimality equation or that the greedy policy is optimal. Both can fail to be true. There are several ways to fix this deficiency. Perhaps the simplest is to solve the linear program in Eq. (37.5) to find ρ^* and then solve another linear program that fixes the gain while minimizing the value function. Let $\tilde{s} \in \mathcal{S}$ and consider the following linear program:

$$\begin{aligned} & \underset{v \in \mathbb{R}^S}{\text{minimize}} \quad \langle v, \mathbf{1} \rangle \\ & \text{subject to } \rho^* + v(s) \geq r_a(s) + \langle P_a(s), v \rangle \text{ for all } s, a \\ & \quad v(\tilde{s}) = 0. \end{aligned} \tag{37.6}$$

The second constraint is crucial in order for the minimum to exist, since otherwise the value function can be arbitrarily small. The next theorem shows that provided \tilde{s} is chosen appropriately, then the solution of Eq. (37.6) satisfies the Bellman optimality equation.

THEOREM 37.3 *Let v be a solution of Eq. (37.6) and assume there exists an optimal policy π^* such that $P_{\pi^*}^*(s, \tilde{s}) > 0$ for all $s \in \mathcal{S}$. Then (ρ^*, v) satisfies the Bellman optimality equation.*

Proof Let $\varepsilon = v + \rho^* - Tv$, which by the first constraint satisfies $\varepsilon \geq 0$. Let π^* be an optimal policy satisfying the requirements of the theorem statement and π be the greedy policy with respect to v . Then

$$P_{\pi^*}^t r_{\pi^*} \leq P_{\pi^*}^t (r_\pi + P_\pi v - P_{\pi^*} v) = P_{\pi^*}^t (\rho^* + v - \varepsilon - P_{\pi^*} v).$$

Hence $\rho^* = \rho^* = \rho^* - P_{\pi^*} \varepsilon$, which means that $P_{\pi^*}^* \varepsilon = 0$ and so $\varepsilon(s) = 0$ for all states s in any recurrence class of π^* . By our assumption on $P_{\pi^*}^*$ we conclude that $\varepsilon(\tilde{s}) = 0$. It follows that $\tilde{v} = v + \varepsilon$ also satisfies the constraints in Eq. (37.6) and by the assumption that v is a solution we conclude that $\varepsilon = 0$. \square

To complete the procedure we need to find a state \tilde{s} that is recurrent under some optimal policy. There is a relatively simple procedure for doing this using the solution to Eq. (37.5), but its analysis depends on the basic theory of duality from the linear programming. Instead we note that one can simply solve Eq. (37.6) for all choices of \tilde{s} and take the first solution that satisfies the Bellman optimality equation.

37.3.1 Efficient computation

The general form of a linear program is an optimization problem of the form

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \langle c, x \rangle \\ & \text{subject to } Ax \geq b, \end{aligned}$$

where $c \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ are parameters of the problem. This general problem can be solved in time that depends polynomially on n and m . When m is very large or infinite these algorithms may become impractical, but nevertheless one can often still solve the optimization problem in time polynomial in n only, provided that the constraints satisfy certain structural properties. Let $\mathcal{K} \subset \mathbb{R}^n$ be convex and consider

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \langle c, x \rangle \\ & \text{subject to } x \in \mathcal{K}. \end{aligned} \tag{37.7}$$

Algorithms for this problem generally have a slightly different flavor because \mathcal{K} may have no corners. Suppose the following are known:

- (a) There exists a known $R > 0$ such that $\mathcal{K} \subset \{x \in \mathbb{R}^n : \|x\| \leq R\}$.
- (b) There exist a separation oracle, which is a function ϕ on \mathbb{R}^n with $\phi(x) = \text{TRUE}$ for $x \in \mathcal{K}$ and otherwise $\phi(x) = u$ and $\langle y, u \rangle > \langle x, u \rangle$ for all $y \in \mathcal{K}$ (see Fig. 37.3).

Given a separation oracle and a bound R on the size of \mathcal{K} the ellipsoid method can solve Eq. (37.7) to ε accuracy in time polynomial in n and $\log(R/\varepsilon)$. The reader can find references to this method at the end of the chapter. The final

step is to give a condition when a separation oracle exists for the convex sets determined by the constraints in Eq. (37.5) and Eq. (37.6). Assuming that

$$\operatorname{argmax}_{a \in \mathcal{A}} (r_a(s) + \langle P_a(s), v \rangle) \quad (37.8)$$

can be solved efficiently, then Algorithm 23 provides a separation oracle. For the specialized case considered later Eq. (37.8) is trivial to compute efficiently.

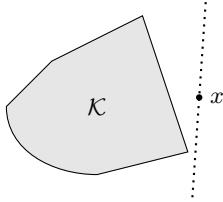


Figure 37.3 Separation oracle returns the normal of a hyperplane that separates x from \mathcal{K} whenever $x \notin \mathcal{K}$. When $x \in \mathcal{K}$ the separation oracle returns TRUE.



In Theorem 37.1 we assumed an exact solution of the Bellman optimality equation, which may not be possible in practice. Fortunately, approximate solutions to the Bellman optimality equation yield approximately optimal greedy policies. Details are in Exercise 37.24.

```

1: function SEPARATIONORACLE( $\rho, v$ )
2:   For each  $s \in \mathcal{S}$  find  $a_s^* \in \operatorname{argmax}_a (r_a(s) + \langle P_a(s), v \rangle)$ 
3:   if  $\rho + v(s) \geq r_{a_s^*}(s) + \langle P_{a_s^*}(s), v \rangle$  for all  $s \in \mathcal{S}$  then
4:     return TRUE
5:   else
6:     Find state  $s$  with  $\rho + v(s) < r_{a_s^*}(s) + \langle P_{a_s^*}(s), v \rangle$ 
7:     Return  $(1, e_s - P_{a_s^*}(s))$ 
8:   end if
9: end function
```

Algorithm 23: Separation oracle.

37.4 Learning in Markov decision processes

When the Markov decision process is unknown the problem of finding an optimal policy is no longer just an optimization problem and the regret is introduced to measure the price of the uncertainty. For simplicity we assume that only the transition matrix is unknown while the reward function is given. This assumption

is not especially restrictive as the case where the rewards are also unknown is easily covered using either a reduction or a simple generalization as we explain in the notes. The regret of a policy π is the deficit of rewards suffered relative to the expected average reward of an optimal policy:

$$\hat{R}_n = n\rho^* - \sum_{t=1}^n r_{A_t}(S_t).$$

The reader will notice we are comparing the nonrandom $n\rho^*$ to the random sum of rewards received by the learner, which was also true in the study of stochastic bandits. The difference is that ρ^* is an asymptotic quantity while for stochastic bandits the analogous quantity was $n\mu^*$. The definition stills makes sense, however, because for MDPs with finite diameter D the optimal expected value over n rounds is at least $n\rho^* - D$ so the difference is negligible (Exercise 37.15). The main result of this chapter is the following.

THEOREM 37.4 *Let $C > 0$ be a sufficiently large universal constant, S and A be positive integers, and $\delta \in (0, 1)$. Then there exists a policy π such that for any MDP $M = (\mathcal{S}, \mathcal{A}, P, r)$ with S states, A actions and rewards from $[0, 1]$, any initial state distribution $\mu \in \mathcal{P}(\mathcal{S})$ and for any horizon $n \geq 1$,*

$$\mathbb{P}\left(\hat{R}_n \geq CD(M)S\sqrt{An \log(nSA/\delta)}\right) \leq \delta.$$

In Exercise 37.17 we ask you to use the assumption that the rewards are bounded to find a choice of $\delta \in (0, 1)$ such that

$$\mathbb{E}[\hat{R}_n] \leq 1 + CD(M)S\sqrt{2An \log(n)}. \quad (37.9)$$

This result is complemented by the following lower bound.

THEOREM 37.5 *Let $S \geq 3$, $A \geq 2$, $D \geq 6 + 2 \log_A S$ and $n \geq DSA$. Then for any policy there exists a Markov decision process with S states, A actions and diameter at most D such that*

$$\mathbb{E}[\hat{R}_n] \geq C\sqrt{DSAn},$$

where $C > 0$ is again a universal constant.

The upper and lower bounds are separated by a factor of at least \sqrt{DS} , which is a considerable gap. Recent work has made progress towards closing this gap as we explain in the notes.

37.5 Upper confidence bounds for reinforcement learning

The algorithm that establishes Theorem 37.4 combines the use of phases with the optimism principle. At the start of each phase the algorithm computes an optimal policy for the statistically plausible MDP with the largest optimal gain. This policy is then implemented until the number of visits to some state/action

pair doubles when a new phase starts and the process begins again. The use of phases is important, not just for computational efficiency. Recalculating the optimistic policy in each round may lead to a dithering behavior in which the algorithm frequently changes its plan and suffers linear regret (Exercise 37.22). We first define confidence sets on the unknown quantity, which in this case is the transition matrix. The confidence sets are centered at the empirical transition probabilities defined by

$$\hat{P}_{t,a}(s, s') = \frac{\sum_{u=1}^t \mathbb{I}\{S_u = s, A_u = a, S_{u+1} = s'\}}{1 \vee T_t(s, a)},$$

where $T_t(s, a) = \sum_{u=1}^t \mathbb{I}\{S_u = s, A_u = a\}$ is the number of times action a was taken in state s . As before we let $\hat{P}_{t,a}(s)$ be the vector whose s' th entry is $\hat{P}_{t,a}(s, s')$. Given a state/action pair s, a define

$$\mathcal{C}_t(s, a) = \left\{ P \in \mathcal{P}(\mathcal{S}) : \|P - \hat{P}_{t-1,a}(s)\|_1 \leq \sqrt{\frac{SL_{t-1}(s, a)}{1 \vee T_{t-1}(s, a)}} \right\}, \quad (37.10)$$

where for $T_t(s, a) > 0$ we set

$$L_t(s, a) = 2 \log \left(\frac{4SAT_t(s, a)(1 + T_t(s, a))}{\delta} \right)$$

and for $T_t(s, a) = 0$ we set $L_t(s, a) = 1$. Note that in this case $\mathcal{C}_{t+1}(s, a) = \mathcal{P}(\mathcal{S})$. Then define confidence set on the space of transition kernels by

$$\mathcal{C}_t = \{P = (P_a(s))_{s,a} : P_a(s) \in \mathcal{C}_t(s, a) \text{ for all } s, a \in \mathcal{S} \times \mathcal{A}\}, \quad (37.11)$$

Clearly $T_t(s, a)$ cannot be larger than the total number of rounds n so

$$L_t(s, a) \leq L = 2 \log \left(\frac{4SAN(n+1)}{\delta} \right). \quad (37.12)$$

The algorithm operates in phases $k = 1, 2, 3, \dots$ with the first phase starting in round $\tau_1 = 1$ and the $(k+1)$ th phase starting in round τ_{k+1} defined inductively by

$$\tau_{k+1} = 1 + \min \{t : T_t(S_t, A_t) \geq 2T_{\tau_k-1}(S_t, A_t)\},$$

which means that the next phase starts once the number of visits to some state at least doubles.

37.5.1 The extended Markov decision process

The confidence set \mathcal{C}_t defines a set of plausible transition probability functions at the start of round t . Since the reward function is known already this corresponds to a set of plausible MDPs. The algorithm plays according to the optimal policy in the plausible MDP with the largest gain. There is some subtlety because the optimal policy is not unique and what is really needed is to find a policy that is greedy with respect to a value function satisfying the Bellman optimality

equation in the plausible MDP with the largest gain. Precisely, at the start of the k th phase the algorithm must find a value function v_k , gain ρ_k and MDP $M_k = (\mathcal{S}, \mathcal{A}, P_k, r)$ with $P_k \in \mathcal{C}_{\tau_k}$ such that

$$\begin{aligned}\rho_k + v_k(s) &= \max_{a \in \mathcal{A}} r_a(s) + \langle P_{k,a}(s), v_k \rangle \text{ for all } s \in \mathcal{S} \text{ and } a \in \mathcal{A}, \\ \rho_k &= \max_{s \in \mathcal{S}} \max_{\pi \in \Pi_{\text{DM}}} \max_{P \in \mathcal{C}_{\tau_k}} \rho_s^\pi(P),\end{aligned}\tag{37.13}$$

where $\rho_s^\pi(P)$ is the gain of deterministic memoryless policy π starting in state s in the MDP with transition probability function P . The algorithm then plays according to π_k defined as the greedy policy with respect to v_k . There is quite a lot hidden in these equations. The gain is only guaranteed to be constant when M_k has a finite diameter, but this may not hold for all plausible MDPs. As it happens, however, solutions to Eq. (37.13) are guaranteed to exist and can be found efficiently. To see why this is true we introduce the **extended Markov decision process** \tilde{M}_k , which has state-space \mathcal{S} and state-dependent action-space $\tilde{\mathcal{A}}_s$ given by

$$\tilde{\mathcal{A}}_s = \{(a, P) : a \in \mathcal{A}, P \in \mathcal{C}_{\tau_k}(s, a)\}.$$

The reward function of the extended MDP is $\tilde{r}_{(a,P)}(s) = r_a(s)$ and the transitions are $\tilde{P}_{a,P}(s) = P_a(s)$. The action-space in the extended MDP allows the agent to choose both $a \in \mathcal{A}$ and a plausible transition vector $P_a(s) \in \mathcal{C}_{\tau_k}(s, a)$. By the definition of the confidence sets, for any pair of states s, s' and action $a \in \mathcal{A}$ there always exists a transition vector $P_a(s) \in \mathcal{C}_{\tau_k}(s, a)$ such that $P_a(s, s') > 0$, which means that \tilde{M}_k is strongly connected. Hence solving the Bellman optimality equation for \tilde{M}_k yields a value function v_k and constant gain $\rho_k \in \mathbb{R}$ that satisfy Eq. (37.13). A minor detail is that the extended action-sets are infinite while the analysis in previous sections only demonstrated existence of solutions to the Bellman optimality equation for finite MDPs. We leave it to the reader to convince themselves that $\mathcal{C}_t(s, a)$ is convex and has finitely many extremal points. Restricting the confidence sets to these points makes the extended MDP finite without changing the optimal policy.

```

1: Input  $\mathcal{S}, \mathcal{A}, r, \delta \in (0, 1)$ 
2:  $t = 0$ 
3: for  $k = 1, 2, \dots$  do
4:    $\tau_k = t + 1$ 
5:   Find  $\pi_k$  as the greedy policy with respect  $v_k$  satisfying Eq. (37.13)
6:   do
7:      $t \leftarrow t + 1$ , observe  $S_t$  and take action  $A_t = \pi_k(S_t)$ 
8:     while  $T_t(S_t, A_t) < 2T_{\tau_k-1}(S_t, A_t)$ 
9:   end for
```

Algorithm 24: UCRL2

37.5.2 Computing the optimistic policy (\dagger)

Here we explain how to efficiently solve the Bellman optimality equation for the extended MDP. The results in Section 37.3 show that the Bellman optimality equation for \tilde{M}_k can be solved efficiently provided that for any value function $v \in \mathbb{R}^S$ the following computation is efficient.

$$\operatorname{argmax}_{a \in \mathcal{A}} \left(r_a(s) + \max_{P \in \mathcal{C}_{\tau_k}(s, a)} \langle P, v \rangle \right). \quad (37.14)$$

The inner optimization is another linear program with S variables and $O(S)$ constraints and can be solved in polynomial time. This procedure is repeated for each $a \in \mathcal{A}$ to solve the whole thing. In fact the inner optimization can be solved more straightforwardly by sorting the entries of v and then allocating P coordinate-by-coordinate to be as large as allowed by the constraints in decreasing order of v . The total computation cost of solving Eq. (37.14) in this way is $O(S(A + \log S))$. Combining this with Algorithm 23 gives the required separation oracle.

The next problem is to find an R such that the set of feasible solutions to the linear programs in Eq. (37.5) and Eq. (37.6) are contained in the set $\{x : \|x\| \leq R\}$. For Eq. (37.5) no such R exists because solutions for the value function remain solutions when translated. To get around this it is necessary to add a constraint that $v \leq b$ and $v \geq -b$ for some carefully chosen $b \in \mathbb{R}$. By Lemma 37.1 the span of the value function satisfying the Bellman optimality equation is at most the diameter. Note that for each pair of states s, s' there exists an action a and $P \in \mathcal{C}_{\tau_k}(s, a)$ such that $P(s, s') \geq \min\{1, \sqrt{n}\}$ so that $D(\tilde{M}_k) \leq \sqrt{n}$. Then we may choose $b = \sqrt{n}$ and $R = \sqrt{1 + nS}$, where we used the fact that the optimal gain is at most 1. Combining this with the tools developed in Section 37.3 shows that the Bellman optimality equation for \tilde{M}_k may be solved using linear programming in polynomial time. Note the additional constraints requires a minor adaptation of the separation oracle, which we leave for the reader.

37.6 Proof of upper bound

The proof is developed in three steps. First we decompose the regret into phases and define a failure event where the confidence intervals fail. In the second step we bound the regret in each phase and in the third step we sum over the phases. Recall that $M = (\mathcal{S}, \mathcal{A}, P, r)$ is the true Markov decision process with diameter $D = D(M)$. The initial state distribution is $\mu \in \mathcal{P}(\mathcal{S})$, which is arbitrary.

Step 1: Failure events and decomposition

Let K be the (random) number of phases and for $k \in [K]$ let $E_k = \{\tau_k, \tau_k + 1, \dots, \tau_{k+1} - 1\}$ be the set of rounds in the k th phase where τ_{K+1} is defined to be $n + 1$. Let $T_{(k)}(s, a)$ be the number of times state/action pair s, a is visited in

the k th phase:

$$T_{(k)}(s, a) = \sum_{t \in E_k} \mathbb{I}\{S_t = s, A_t = a\} .$$

Define F as the failure event that $P \notin \mathcal{C}_{\tau_k}$ for some $k \in [K]$. The first lemma shows that F has lower probability:

LEMMA 37.2 $\mathbb{P}(F) \leq \delta/2$.

The proof is based on a concentration inequality derived for categorical distributions and is left for Exercise 37.8. When F does not hold the true transition kernel is in \mathcal{C}_{τ_k} for all k , which means that $\rho^* \leq \rho_k$ and

$$\hat{R}_n = \sum_{t=1}^n (\rho^* - r_{A_t}(S_t)) \leq \sum_{k=1}^K \underbrace{\sum_{t \in E_k} (\rho_k - r_{A_t}(S_t))}_{\tilde{R}_k} .$$

In the next step we bound \tilde{R}_k under the assumption that F does not hold.

Step 2: Bounding the regret in each phase

Assume that F does not occur and fix $k \in [K]$. Recall that v_k is a value function satisfying the Bellman optimality equation in the optimistic MDP M_k and ρ_k is its gain. Hence

$$\rho_k = r_{\pi_k}(s) - v_k(s) + P_{k,\pi_k}(s)^\top v_k \quad \text{for all } s \in \mathcal{S}. \quad (37.15)$$

As noted earlier, solutions to the Bellman optimality equation remain solutions when translated so we may choose v_k such that $\|v_k\|_\infty \leq \text{span}(v_k)/2$, which means that

$$\|v_k\|_\infty \leq \frac{1}{2} \text{span}(v_k) \leq \frac{D}{2}, \quad (37.16)$$

where the second inequality follows from Lemma 37.1 and the fact that when F does not hold the diameter of M_k is at most D . By the definition of the policy we have $A_t = \pi_k(S_t)$ for $t \in E_k$, which implies that

$$\rho_k = r_{A_t}(S_t) - v_k(S_t) + P_{k,A_t}(S_t)^\top v_k \quad \text{for all } t \in E_k.$$

Rearranging and substituting yields

$$\begin{aligned} \tilde{R}_k &= \sum_{t \in E_k} (-v_k(S_t) + P_{k,A_t}(S_t)^\top v_k) \\ &= \sum_{t \in E_k} (-v_k(S_t) + P_{A_t}(S_t)^\top v_k) + \sum_{t \in E_k} (P_{k,A_t}(S_t) - P_{A_t}(S_t))^\top v_k \\ &\leq \underbrace{\sum_{t \in E_k} (-v_k(S_t) + P_{A_t}(S_t)^\top v_k)}_{(A)} + \underbrace{\frac{D}{2} \sum_{t \in E_k} \|P_{k,A_t}(S_t) - P_{A_t}(S_t)\|_1}_{(B)}, \end{aligned} \quad (37.17)$$

where the inequality follows from Hölder's inequality and Eq. (37.16). Let $\mathbb{E}_t[\cdot]$ denote the conditional expectation with respect to \mathbb{P} conditioned on $\sigma(S_1, A_1, \dots, S_{t-1}, A_{t-1}, S_t)$. To bound (A) we reorder the terms and use the fact that $\text{span}(v_k) \leq D$ on the event F^c .

$$\begin{aligned} (A) &= \sum_{t \in E_k} (v_k(S_{t+1}) - v_k(S_t) + P_{A_t}(S_t)^\top v_k - v_k(S_{t+1})) \\ &= v_k(S_{\tau_{k+1}}) - v_k(S_{\tau_k}) + \sum_{t \in E_k} (P_{A_t}(S_t)^\top v_k - v_k(S_{t+1})) \\ &\leq D + \sum_{t \in E_k} (\mathbb{E}_t[v_k(S_{t+1})] - v_k(S_{t+1})) , \end{aligned}$$

where the second equality used that $\max E_k = \tau_{k+1} - 1$ and $\min E_k = \tau_k$. We leave this here for now and move on to term (B) in Eq. (37.17). The definition of the confidence intervals and the assumption that F does not occur shows that

$$(B) \leq \frac{D\sqrt{LS}}{2} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{T_{(k)}(s,a)}{\sqrt{1 \vee T_{\tau_k-1}(s,a)}} .$$

Combining the bounds (A) and (B) yields

$$\tilde{R}_k \leq D + \sum_{t \in E_k} (\mathbb{E}_t[v_k(S_{t+1})] - v_k(S_{t+1})) + \frac{D\sqrt{LS}}{2} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{T_{(k)}(s,a)}{\sqrt{1 \vee T_{\tau_k-1}(s,a)}} .$$

Step 3: Bounding the number of phases and summing

Let K_t be the phase in round t so that $t \in E_{K_t}$. By the work in the previous two steps, if F does not occur then

$$\begin{aligned} \hat{R}_n &\leq \sum_{k=1}^K \tilde{R}_k \leq KD + \sum_{t=1}^n (\mathbb{E}_t[v_{K_t}(S_{t+1})] - v_{K_t}(S_{t+1})) \\ &\quad + \frac{D\sqrt{LS}}{2} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \sum_{k=1}^K \frac{T_{(k)}(s,a)}{\sqrt{1 \vee T_{\tau_k-1}(s,a)}} . \end{aligned}$$

The first sum is bounded using a version of Hoeffding–Azuma (Exercise 20.6):

$$\mathbb{P} \left(F^c \text{ and } \sum_{t=1}^n (\mathbb{E}_t[v_{K_t}(S_{t+1})] - v_{K_t}(S_{t+1})) \geq D \sqrt{\frac{n \log(2/\delta)}{2}} \right) \leq \frac{\delta}{2} .$$

For the second term we note that $T_{(k)}(s,a)/\sqrt{1 \vee T_{\tau_k-1}(s,a)}$ cannot be large too often. A continuous approximation often provides intuition for the correct form. Recalling the thousands of integrals you did at school, for any differentiable $f : [0, \infty) \rightarrow \mathbb{R}$ we have

$$\int_0^K \frac{f'(k)}{\sqrt{f(k)}} dk = 2\sqrt{f(K)} - 2\sqrt{f(0)} .$$

Here we are thinking of $f(k)$ as the continuous approximation of $T_{\tau_k-1}(s, a)$ and its derivative as $T_{(k)}(s, a)$. In Exercise 37.20 we ask you to make this argument rigorous by showing that

$$\sum_{k=1}^K \frac{T_{(k)}(s, a)}{\sqrt{T_{\tau_k-1}(s, a)}} \leq (\sqrt{2} + 1) \sqrt{T_n(s, a)}.$$

Then by Cauchy-Schwartz and the fact that $\sum_{s,a \in \mathcal{S} \times \mathcal{A}} T_n(s, a) = n$,

$$\sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sqrt{T_n(s, a)} \leq \sqrt{SA n}.$$

It remains to bound the number of phases. A new phase starts when the visit count for some state/action pair doubles. Hence K cannot be more than the number of times the counters double in total for each of the states. It is easy to see that $1 + \log_2 T_n(s, a)$ gives an upper bound on how many times the counter for this pair may double (the constant 1 is there to account for the counter changing from zero to one). Thus $K \leq K' = \sum_{s,a} 1 + \log_2 T_n(s, a)$. Noting that $0 \leq T_n(s, a)$ and $\sum_{s,a} T_n(s, a) = n$ and relaxing $T_n(s, a)$ to take real values we find that the value of K' is the largest when $T_n(s, a) = n/(SA)$, which shows that

$$K \leq SA \left(1 + \log_2 \left(\frac{n}{SA} \right) \right).$$

Putting everything together gives the desired result.

37.7 Proof of lower bound

The lower bound is proven by crafting a difficult MDP that models a bandit with approximately SA arms. This a cumbersome endeavour, but intuitively straightforward and the explanations that follow should be made clear in Fig. 37.4. Given S and A the first step is to construct a tree of minimum depth with at most A children for each node using exactly $S - 2$ states. The root of the tree is denoted by s_o and transitions within the tree are deterministic, so in any given node the learner can simply select which child to transition to. Let L be the number of leaves and label these states s_1, \dots, s_L . The last two states are s_g and s_b ('good' and 'bad' respectively). For each $i \in [L]$ the learner can take any action $a \in \mathcal{A}$ and transitions to either the good state or the bad state according to

$$P_a(s_i, s_g) = \frac{1}{2} + \varepsilon(a, i) \quad \text{and} \quad P_a(s_i, s_b) = \frac{1}{2} - \varepsilon(a, i).$$

The function ε will be chosen so that $\varepsilon(a, i) = 0$ for all (a, i) pairs except one. For this special state/action pair we let $\varepsilon(a, i) = \Delta$ for appropriately tuned $\Delta > 0$. The good state and the bad state have the same transitions for all actions.

$$\begin{aligned} P_a(s_g, s_g) &= 1 - \delta & P_a(s_g, s_o) &= \delta \\ P_a(s_b, s_b) &= 1 - \delta & P_a(s_b, s_o) &= \delta, \end{aligned}$$

where $\delta = 4/D$, which under the assumptions of the theorem is guaranteed to be in $(0, 1]$ and is chosen to ensure the diameter of the described MDP is at most D . The reward function is $r_a(s) = 1$ if $s = s_g$ and $r_a(s) = 0$ otherwise.

The connection to finite-armed bandits is straightforward. Each time the learner arrives in state s_o it selects which leaf to visit and then chooses an action from that leaf. This corresponds to choosing one of $K = LA = \Omega(\text{SA})$ meta actions. The optimal policy is to select the meta action with the largest probability of transitioning to the good state. The choice of δ means the learner expects to stay in the good/bad state for approximately D rounds, which also makes the diameter of this MDP about D . All up this means the learner expects to make about n/D decisions and the rewards are roughly in $[0, D]$ so we should expect the regret to be $\Omega(D\sqrt{n/DK}) = \Omega(\sqrt{nDSA})$.

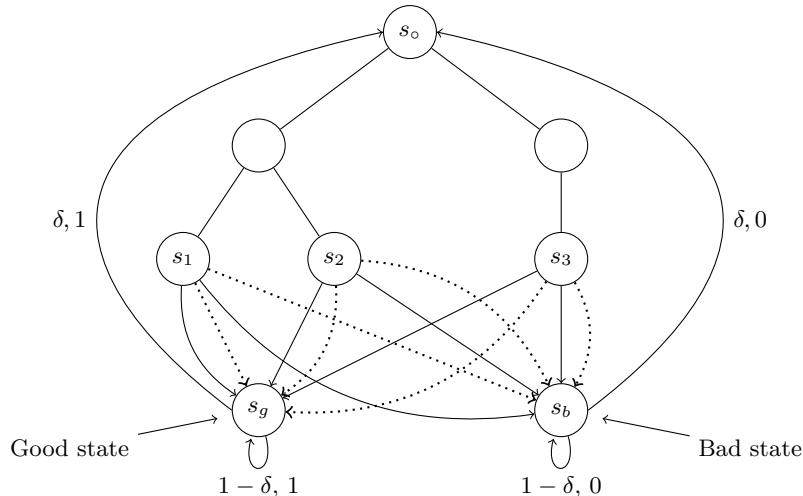


Figure 37.4 Lower bound construction for $A = 2$ and $S = 8$. The resulting MDP is roughly equivalent to a bandit with six actions

One could almost claim victory here and not bother with the proof. As usual, however, there are some technical difficulties, which in this case arise because the number of visits to the decision state s_o is a random quantity. For this reason we give the proof, leaving as exercises the parts that are both obvious and annoying.

Proof of Theorem 37.5 The proof follows the path suggested in Exercise 15.1. We break things up into two steps. Throughout we fix an arbitrary policy π .

Step 1: Notation and facts about the MDP

Let d be the depth of the tree in the MDP construction and L the number of leaves and $K = LA$ ($d = 3$ and $L = 3$ in Fig. 37.4). This leaves K state-action pairs that potentially lead to either s_g or s_b . Let M_0 be the MDP with $\varepsilon(s, a) = 0$ for all relevant state-action pairs s, a and M_k be the MDP with $\varepsilon(s, a) = \Delta$ for

the k th state/action pair with the state on the fringe of the tree, ordered in some arbitrary way. Define stopping time τ by

$$\tau = n \wedge \min \left\{ t : \sum_{s=1}^t \mathbb{I}\{S_t = s_0\} \geq \frac{n}{D} \right\},$$

which is the first round when the number of visits to state s_0 is at least n/D . Next let T_k be the number of visits to state-action pair k until stopping time τ and $T_\sigma = \sum_{k=1}^K T_k$. Let \mathbb{P}_k be the law of T_1, \dots, T_K induced by the interaction of π and M_k and $\mathbb{E}_k[\cdot]$ the expectation with respect to \mathbb{P}_k . None of the following claims are surprising, but are tiresome to prove. They are listed in increasing order of difficulty and left to the reader in Exercise 37.25.

CLAIM 37.1 *For all $k \in [K]$ the diameter is bounded by $D(M_k) \leq D$.*

CLAIM 37.2 *There exist universal constants $0 < c_1 < c_2 < \infty$ such that*

$$D\mathbb{E}_0[T_\sigma]/n \in [c_1, c_2].$$

CLAIM 37.3 *Let R_{nk} be the expected regret of policy π in MDP M_k over n rounds. There exists a universal constant $c_3 > 0$ such that*

$$R_{nk} \geq c_3 \Delta D (\mathbb{E}_k[T_\sigma - T_k]).$$

Step 2: Bounding the regret

The relative entropy between \mathbb{P}_0 and \mathbb{P}_k is easily calculated by noticing that M_0 and M_k only differ when state-action pair k is visited and then using Lemma 15.1 and the entropy inequalities Eq. (14.11) and the assumption that $\Delta \leq 1/2$,

$$D(\mathbb{P}_0, \mathbb{P}_k) = \mathbb{E}[T_k]d(1/2, 1/2 + \Delta) \leq 4\Delta^2\mathbb{E}_0[T_k],$$

where $d(p, q)$ is the relative entropy between Bernoulli distributions with biases p and q respectively. Using the fact that $T_\sigma \leq n/D$ and Pinsker's inequality (Eq. (14.8)),

$$\mathbb{E}_k[T_\sigma - T_k] \geq \mathbb{E}_0[T_\sigma - T_k] - \frac{n}{D} \sqrt{\frac{D(\mathbb{P}_0, \mathbb{P}_k)}{2}} \geq \mathbb{E}_0[T_\sigma - T_k] - \frac{n\Delta}{D} \sqrt{2\mathbb{E}_0[T_k]}.$$

Summing over k and applying Cauchy-Schwartz yields

$$\begin{aligned} \sum_{k=1}^K \mathbb{E}_k[T_\sigma - T_k] &\geq \sum_{k=1}^K \mathbb{E}_0[T_\sigma - T_k] - \frac{n\Delta}{D} \sum_{k=1}^K \sqrt{2\mathbb{E}_0[T_k]} \\ &\geq (K-1)\mathbb{E}_0[T_\sigma] - \frac{n\Delta}{D} \sqrt{2K\mathbb{E}_0[T_\sigma]} \\ &\geq \frac{c_1 n(K-1)}{D} - \frac{n\Delta}{D} \sqrt{\frac{2c_2 n K}{D}} \\ &\geq \frac{c_1 n(K-1)}{2D}, \end{aligned} \tag{37.18}$$

where the last inequality follows by choosing

$$\Delta = \frac{c_1(K-1)}{2} \sqrt{\frac{D}{2c_2nK}}.$$

By Eq. (37.18) there exists a $k \in [K]$ such that

$$\mathbb{E}_k[T_\sigma - T_k] \geq \frac{c_1 n(K-1)}{2DK}.$$

Then for last step apply Claim 37.3 to show that

$$R_{nk} \geq c_3 D \Delta \mathbb{E}_k[T_\sigma - T_k] \geq \frac{c_1^2 c_3 n(K-1)^2}{4K} \sqrt{\frac{D}{2c_2nK}}.$$

Naive bounding and simplification concludes the result. \square

37.8 Notes

- 1 ‘Operator’ it is just a fancy word for ‘function’. The Bellman operator is a function from the space of value functions to itself. Operators are usually denoted by capital letters and brackets are omitted in their application so that Tv is shorthand for $T(v)$. It is not a requirement of the definition, but operators are usually defined on spaces of functions and preserve certain structures of the space.
- 2 MDPs in applications can have millions (or “Billions and Billions”) of states, which should make the reader worried that the bound in Theorem 37.4 could be extremely large. The takeaway should be that learning in large MDPs without additional assumptions is hard, as attested by the lower bound in Theorem 37.5.
- 3 The key to choosing the state space is that the state must be observable and sufficiently informative that the Markov property is satisfied. Blowing up the size of the state space may help to increase the fidelity of the approximation (the entire history always works), but will almost always slow down learning.
- 4 A state $s \in \mathcal{S}$ is **absorbing** if $P_a(s, s) = 1$ for all $a \in \mathcal{A}$. An MDP is **episodic** if there exists an absorbing state that is reached almost surely by any policy. The average reward criterion is meaningless in episodic MDPs because all policies are optimal. In this case the usual objective is to maximize the expected reward until the absorbing state is reached without limits or normalization, sometimes with discounting. An MDP is **finite-horizon** if it is episodic and the absorbing state is always reached after some fixed number of rounds. The learning community studies these in the same way as bandits, where in each ‘round’ the learner interacts with the MDP from some starting state until the absorbing state is reached. The simplification of the setting eases the analysis and preserves most of the intuition from the general setting.

- 5 A **partially observable MDP** is a generalization where the learner does not observe the underlying state. Instead they receive an observation that is a (possibly random) function of the state. Given a fixed (known) initial state distribution, any POMDP can be mapped into an MDP at the price of enlarging the state space. A simple way to achieve this is to let the new state be the space of all histories. However, this actually loses some information. This is a subtle issue that is worth explaining: In an MDP actions do not carry information: In an MDP we can give names specific to the state to every action without changing the information structure available to the decision maker. If we did the same in a POMDP, we would give away information about the state. Hence, in a POMDP the fact that actions are shared across the states has information, which would be lost if we learn using histories. A better way is to use a sufficient statistic for the hidden state as the state. A natural choice is the posterior distribution over the hidden state given the interaction history, which is called the **belief space**. While the value function over the belief space has some nice structure, in general even computing the optimal policy is hard [Papadimitriou and Tsitsiklis, 1987].
- 6 In applications where the asymptotic nature of gain optimality is unacceptable there are more sensitive criteria. A memoryless policy π^* is **bias optimal** if it is gain optimal and $v_{\pi^*} \geq v_\pi$ for all memoryless policies π . Even more sensitive criteria also exist. Some keywords to search for are **Blackwell optimality** and **n -discount** optimality.
- 7 The **Cesàro sum** of a real-valued sequence $(a_n)_n$ is the asymptotic average of the partial sums. Let $s_n = a_0 + \dots + a_{n-1}$ be the n th partial sum. The Cesàro sum of this sequence is $A = \lim_{n \rightarrow \infty} \frac{1}{n}(s_1 + \dots + s_n)$ when this limit exists. The idea is that Cesàro summation smoothens out periodicity and thus increases the range of summable sequences. For example, the alternating sequence $(+1, -1, +1, -1, \dots)$ is Cesàro summable and its Cesàro sum is easily seen to be $1/2$, while it is clearly not summable in the normal sense. If the sequence is summable, its sum and its Cesàro sum are equal. The differential value of a policy is defined as a Cesàro sum so that it is well-defined even if the underlying Markov chain has periodic states.
- 8 For $\gamma \in (0, 1)$ the γ discounted sum of sequence $(a_n)_n$ is $A_\gamma = \sum_{n=0}^{\infty} \gamma^n a_n$. An elementary argument shows that for any $s \in \mathbb{R}$ and sequence $(a_n)_n$ for which A_α is well-defined, $A_\gamma = \frac{s}{1-\gamma} + (1-\gamma) \sum_{n=1}^{\infty} \gamma^{n-1} (s_n - is)$. When s is the Cesàro sum A of $(a_n)_n$ it is not hard to see that $|\sum_{i=1}^{\infty} \gamma^{i-1} (s_i - iA)| = O(1/(1-\gamma))$ and thus $(1-\gamma)A_\gamma - A = O((1-\gamma))$ as $\gamma \rightarrow 1$. The approach of approximating Cesàro sums through discounted sums with the discount factor γ approaching one is called the **vanishing discount approach**.
- 9 We mentioned enumeration, value iteration and policy iteration as other methods for computing optimal policies. Enumeration just means enumerating all deterministic memoryless policies and selecting the one with the highest gain. This is obviously too expensive. **Policy iteration** is an iterative process

that starts with a policy π_0 . In each round the algorithm computes π_{k+1} from π_k by computing v_{π_k} and then choosing π_{k+1} to be the greedy policy with respect to v_{π_k} . In general this method may not converge to an optimal policy, but by slightly modifying the update process one can prove convergence. For more details see the Chapter 4 of Volume 2 of the book by [Bertsekas \[2012\]](#).

Value iteration works by choosing an arbitrary value function v_0 and then inductively defining $v_{k+1} = T v_k$ where $(T v)(s) = \max_{a \in \mathcal{A}} r_a(s) + \langle P_a(s), v \rangle$ is the Bellman operator. Under certain technical conditions one can prove that the greedy policy with respect to v_k converges to an optimal policy. Note that $v_{k+1} = \Omega(k)$, which can be a problem numerically. A simple idea is to let $v_{k+1} = T v_k - \delta_k$ where $\delta_k = \max_{s \in \mathcal{S}} v_k(s)$. Since the greedy policy is the same for v and $v + c\mathbf{1}$ this does not change the mathematics, but improves the numerical situation. The aforementioned book by [Bertsekas](#) is again a good source for more details. Unfortunately none of these algorithms have known polynomial time guarantees on the computation complexity of finding an optimal policy without stronger assumptions than we would like. In practice, however, both value and policy iteration work quite well, while the ellipsoid method for solving linear programs should be avoided at all costs.

- 10 One can modify the concept of regret to allow for MDPs that have traps, allowing for finite MDPs with infinite diameter. The idea is as follows: In any finite MDP there exists finitely many disjoint classes of states (what these classes are depends only on the MDP structure) so that each class is a trap in the sense that no policy can escape from it once entered. Now, rule out all those policies that have linear regret in strongly connected MDPs as a reasonable learner should achieve sublinear regret in such MDPs. What remains are policies that will necessarily get trapped in any MDP that is not strongly connected. For such MDPs, the regret is redefined by ‘restarting the clock’ at the time when the policy gets trapped. For details, see Exercise 37.19, where you are also asked to show a policy that achieves sublinear regret in any finite MDP.
- 11 The assumption that the reward function is known can be relaxed without difficulty. It is left as an exercise to figure out how to modify algorithm and analysis to the case when r is unknown and reward observed in round t is bounded in $[0, 1]$ and has conditional mean $r_{A_t}(S_t)$.
- 12 Although it has not been done yet in this setting, the path to removing the spurious \sqrt{S} from the bound is to avoid the application of Cauchy-Schwartz in Eq. (37.17). Instead one should define confidence intervals directly on $\langle \hat{P}_k - P, v_k \rangle$, where the dependence on the state and action has been omitted. Of course this requires one to modify the algorithm. At first sight it seems that one could apply Hoeffding’s bound directly to the inner product, but there is a subtle problem that has spoiled a number of attempts. The problem is that v_k and \hat{P}_k are not independent. This non-independence is unfortunately quite pernicious and appears from many angles. We advise extreme caution (some references for guidance are given at in the bibliographic remarks).

37.9 Bibliographical remarks

The study of sequential decision making has a long history and we recommend the introduction of the book by [Puterman \[2009\]](#) as a good starting point. One of the main architects in modern times is Richard Bellman, who wrote an influential book [[Bellman, 1954](#)]. Bellman had an interesting life, working at Los Alamos near the end of the war and later at RAND. Besides ‘dynamic programming’ he also coined the term ‘curse of dimensionality’ which, although it is not his fault, curses us still today. His autobiography is so entertaining that reading it slowed the writing of this chapter: ‘The Eye of the Hurricane’ [[Bellman, 1984](#)]. As a curiosity, Bellman knew about bandit problems after accidentally encountering a paper by [Thompson \[1935\]](#). For the tidbit see page 260 of the aforementioned biography.



Richard Bellman

Markov decision processes are studied by multiple research communities, including control, operations research and artificial intelligence. The two-volume book by [Bertsekas \[2012\]](#) provides a thorough and formal introduction to the basics. The perspective is quite interdisciplinary, but with a slight (good) bias towards the control literature. The perspective of an operations researcher is most precisely conveyed in the comprehensive book by [Puterman \[2009\]](#). A very readable shorter introductory book is by [Ross \[1983\]](#). [Arapostathis et al. \[1993\]](#) surveyed existing analytical results (existence, uniqueness of optimal policies, validity of the Bellman optimality equation) for average-reward MDPs with an emphasis on continuous state and action space models. The online lecture notes of [Kallenberg \[2016\]](#) are a recent comprehensive alternate account for the theory of discrete MDPs. There are many texts on linear/convex optimization and the ellipsoid method. The introductory book on linear optimization by [Bertsimas and Tsitsiklis \[1997\]](#) is a pleasant read while the ellipsoid method is explained in detail by [Grötschel et al. \[2012\]](#).

The problem considered in this chapter is part of a broader field called reinforcement learning (RL), which has recently seen a surge of interest. The books by [Sutton and Barto \[1998\]](#) and [Bertsekas and Tsitsiklis \[1996\]](#) describe the foundations. The first book provides an intuitive introduction aimed at computer scientists, while the second book focuses on the theoretical results of the fundamental algorithms. A book by one of the present authors focuses on cataloging the range of learning problems encountered in reinforcement learning and summarizing the basic ideas and algorithms [[Szepesvári, 2010](#)].

The UCRL algorithm and the upper and lower regret analysis is due to [Auer et al. \[2009, 2010\]](#). Our proofs differ in minor ways. A more significant difference is that these works used value iteration for finding the optimistic policy and hence

cannot provide polynomial time computation guarantees. In practice this may be preferable to linear programming anyway.

The number of rigorous results for bounding the regret of various algorithms is limited. One idea is to replace the optimistic approach with Thompson sampling, which was first adapted to reinforcement learning by [Strens \[2000\]](#) under the name PSRL (posterior sampling reinforcement learning). [Agrawal and Jia \[2017\]](#) recently made an attempt to improve the dependence of the regret on the state-space. The proof is not quite correct, however, and at the time of writing the holes of not yet been patched. [Azar et al. \[2017\]](#) also improve upon the UCRL2 bound, but for finite-horizon episodic problems where they derive an optimistic algorithm with regret $\tilde{O}(\sqrt{HSAn})$, which after adapting UCRL to the episodic setting improves on its regret by a factor of \sqrt{SH} . The main innovation is to use Freedman's Bernstein-style inequality for computing bonuses directly while computing action values using backwards induction from the end of the episode rather than keeping confidence estimates for the transition probabilities. An issue with both of these improvements is that lower-order terms in the bounds mean they only hold for large n . It remains to be seen if these terms arise from the analysis or if the algorithms need modification.

[Tewari and Bartlett \[2008\]](#) use an optimistic version of linear programming to obtain finite-time logarithmic bounds with suboptimal instance dependent constants. Note this paper mistakenly drops some constants from the confidence intervals, which after fixing would make the constants even worse. Similar results are also available for UCRL2 [\[Auer and Ortner, 2007\]](#). [Burnetas and Katehakis \[1997\]](#) prove asymptotic guarantees with optimal constants, but with the crucial assumption that the support of the next-state distributions $P_a(s)$ are known. [Lai and Graves \[1997\]](#) also consider asymptotic optimality. However, they consider general state spaces where the set of transition probabilities is smoothly parameterized with a known parameterization, but under the weakened goal of competing with the best of finitely many memoryless policies given to the learner as black-boxes.

Finite-time regret for large state and action space MDPs under additional structural assumptions are also considered by [Abbasi-Yadkori and Szepesvári \[2011\]](#), [Abbasi-Yadkori \[2012\]](#), [Ortner and Ryabko \[2012\]](#). [Abbasi-Yadkori and Szepesvári \[2011\]](#) and [Abbasi-Yadkori \[2012\]](#) give algorithms with $O(\sqrt{n})$ regret for linearly parameterized MDP problems with quadratic cost (linear quadratic regulation, or LQR), while [Ortner and Ryabko \[2012\]](#) gives $O(n^{(2d+1)/(2d+2)})$ regret bounds under a Lipschitz assumption, where d is the dimensionality of the state space. The algorithms in these works are not guaranteed to be computationally efficient because they rely on optimistic policies. In theory, this could be addressed by Thompson sampling, which is considered by [Abeille and Lazaric \[2017b\]](#) who obtain partial results for the LQR setting. Thompson sampling has also been studied in the Bayesian framework by [Osband et al. \[2013\]](#), [Abbasi-Yadkori and Szepesvári \[2015\]](#), [Osband and Roy \[2017\]](#), [Theocharous et al. \[2017\]](#), of which [Abbasi-Yadkori and Szepesvári \[2015\]](#) and [Theocharous et al.](#)

[2017] consider general parametrizations, while the other papers are concerned with finite state/action MDPs. Learning in MDPs has also been studied in the Probability Approximately Correct (PAC) framework introduced by Kearns and Singh [2002] where the objective is to design policies for which the number of badly suboptimal actions is small with high probability. The focus of these papers is on the discounted reward setting rather than average reward. The algorithms are again built on the optimism principle. Algorithms that are known to be PAC-MDP include R-max Brafman and Tennenholtz [2003], Kakade [2003], MBIE Strehl and Littman [2005, 2008], Delayed Q-learning Strehl et al. [2006], the optimistic-initialization-based algorithm of Szita and Lörincz [2009], MorMax by Szita and Szepesvári [2010], and an adaptation of UCRL by Lattimore and Hutter [2012], which they call UCRL γ . The latter work presents optimal results (matching upper and lower bounds) for the case when the transition structure is sparse, while the optimal dependence on the number of state/action pairs is achieved by Delayed Q-learning and Mormax [Strehl et al., 2006, Szita and Szepesvári, 2010], though the Mormax bound is better in its dependency on the discount factor. The idea to incorporate the uncertainty in the transitions into the action-space to solve the optimistic optimization problem appeared in the analysis of MBIE [Strehl and Littman, 2008]. A hybrid between stochastic and adversarial settings is when the reward sequence is chosen by an adversary, while transitions are stochastic. This problem has been introduced by Even-Dar et al. [2004]. State-of-the-art results for the bandit case are due to Neu et al. [2014], where the reader can also find further pointers to the literature. The case when both the rewards and the transition probability distributions are also adversarially chosen in various cases by [Abbasi-Yadkori et al., 2013].

37.10 Exercises

37.1 Let $M = (\mathcal{S}, \mathcal{A}, P)$ be a finite **controlled Markov environment**, which is a finite Markov decision process without the reward function. Let π be an arbitrary policy for this environment, i.e., $\pi : \cup_{t=0}^{\infty} (\mathcal{S} \times \mathcal{A})^t \times \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ such that for any $t \geq 0$, $h_t \in (\mathcal{S} \times \mathcal{A})^t \times \mathcal{S}$, $\sum_{a \in \mathcal{A}} \pi(h_t, a) = 1$, and fix a distribution $\mu \in \mathcal{P}(\mathcal{S})$ in an arbitrary manner. Show that there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an infinite sequence $(S_1, A_1, S_2, A_2, \dots)$ of random elements on it such that for $t \in \mathbb{N}$, S_t is \mathcal{S} -valued, A_t is \mathcal{A} -valued, and for any $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$ and $t \in \mathbb{N}$,

- (a) $\mathbb{P}(S_1 = s) = \mu(s);$
- (b) $\mathbb{P}(S_{t+1} = s' | H_t, A_t) = P_{A_t}(S_t, s');$
- (c) $\mathbb{P}(A_t = a | H_t) = \pi(H_t, a),$

where $H_t = (S_1, A_1, \dots, S_{t-1}, A_{t-1}, S_t)$.



Use Theorem 3.3.

37.2 Let $M = (\mathcal{S}, \mathcal{A}, P)$ be a finite controlled Markov environment, π be an arbitrary policy and $\mu \in \mathcal{P}(\mathcal{S})$ an arbitrary initial state distribution.

(a) Show there exists a Markov policy π' such that

$$\mathbb{P}_\mu^\pi(S_t = s, A_t = a) = \mathbb{P}_\mu^{\pi'}(S_t = s, A_t = a).$$

for all $t \geq 1$ and $s, a \in \mathcal{S} \times \mathcal{A}$.

(b) Conclude that for any policy π there exists Markov policies π', π'' such that for any $s \in \mathcal{S}$, $\bar{\rho}_s^\pi = \bar{\rho}_s^{\pi'}$ and $\underline{\rho}_s^\pi = \underline{\rho}_s^{\pi''}$.



Define π' in an inductive manner by first considering $t = 1$, then $t = 2$ and so-on. [Puterman, 2009, Thm. 5.5.1] proves this result and credits Strauch [1966].

37.3 Let P be some transition structure over some finite state space \mathcal{S} and some finite action space \mathcal{A} . Show that the expected travel time between two states s, s' of \mathcal{S} is minimized by a deterministic policy.



Let $\tau^*(s, s')$ be the best expected travel time between some arbitrary pairs of states; for $s = s'$ we define the best travel time to be zero. Show that this satisfies the fixed point equation

$$\tau^*(s, s') = \begin{cases} 0, & \text{if } s = s'; \\ 1 + \min_a \sum_{s''} P_a(s, s'') \tau^*(s'', s'), & \text{otherwise.} \end{cases}$$

37.4 Let M be an MDP. Prove that $D(M) < \infty$ is equivalent to M being strongly connected.

37.5 Let $M = (\mathcal{S}, \mathcal{A}, P, r)$ be any MDP. Show that $D(M) \geq \log_A(S) - 3$.



Denote by $d^*(s, s')$ the minimum expected time it takes to reach state s' when starting from state s . The definition of d^* can be extended to arbitrary initial distributions μ_0 over states and sets $U \subset \mathcal{S}$ of target states: $d^*(\mu_0, U) = \sum_s \mu_0(s) \sum_{s' \in U} d^*(s, s')$. Prove by induction on the size of U that

$$d^*(\mu_0, U) \geq \min \left\{ \sum_{k \geq 0} k n_k \mid 0 \leq n_k \leq A^k, k \geq 0, \sum_{k \geq 0} n_k = |U| \right\} \quad (37.19)$$

and then conclude that the proposition holds by choosing $U = \mathcal{S}$ [Auer et al., 2010, Cor. 15].

37.6 Let e_i be the i th element of the standard Euclidean basis and π be a memoryless policy. Show that $e_i^\top P_\pi^t e_j$ is the probability of arriving in state j from state i in t rounds using policy π .

37.7 Let M be a finite MDP and π a memoryless policy. Prove that for any $i \in \mathcal{S}$ the expected cumulative reward collected by policy π in M is $e_i^\top \sum_{t=1}^n P_\pi^t r_\pi$.

37.8 Prove Lemma 37.2.



Use the result of Exercise 5.19 and apply a union bound over all state/action pairs and the number of samples. Use the Markov property to argue that the independence assumption in Exercise 5.19 is not problematic.

37.9 Let P be any $S \times S$ right stochastic matrix. Show that the following hold:

- (a) $A_n = \frac{1}{n} \sum_{t=0}^{n-1} P^t$ is right stochastic.
- (b) $A_n + \frac{1}{n}(P^n - I) = A_n P = P A_n$.
- (c) $P^* = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=0}^{n-1} P^t$ exists.
- (d) $P^* P = P P^* = P^* P^* = P^*$.
- (e) The matrix $H = (I - P + P^*)^{-1}$ is well-defined.
- (f) Let $D = H - P^*$. Then $D = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \sum_{k=0}^{i-1} (P^k - P^*)$.
- (g) Let $r \in \mathbb{R}^S$ and $\rho = P^* r$. Then $v = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \sum_{k=0}^{i-1} P^k (r - \rho)$ is well-defined.
- (h) With the notation of the previous part, $v + \rho = r + Pv$.



Note that the first four parts of this exercise are the same as in Chapter 36. For Parts (c) and (d) you will likely find it useful that the space of right stochastic matrices is compact. Then show that all cluster points of (A_n) are the same.



The previous exercise implies that the gain and the differential value function of *any* memoryless policy in *any* MDP is well-defined. The matrix H is called the **fundamental matrix** and D is called the **deviation matrix**.

37.10 Let $\gamma \in (0, 1)$ and define operator $T_\gamma : \mathbb{R}^S \rightarrow \mathbb{R}^S$ by

$$(T_\gamma v)(s) = \max_{a \in \mathcal{A}} r_a(s) + \gamma \langle P_a(s), v \rangle.$$

- (a) Prove that T_γ is a contraction with respect to the supremum norm:

$$\|T_\gamma v - T_\gamma w\|_\infty \leq \gamma \|v - w\|_\infty \text{ for any } v, w \in \mathbb{R}^S.$$

- (b) Prove there exists a $v \in \mathbb{R}^S$ such that $T_\gamma v = v$.
- (c) Let π be the greedy policy with respect to v . Show $v = r_\pi + \gamma P_\pi v$.
- (d) Prove that $v = (I - \gamma P_\pi)^{-1} r$.

37.11 Recall that $H = (I - P + P^*)^{-1} - P^*$ and let $P_\gamma^* = (1 - \gamma)(I - \gamma P)^{-1}$. Show that

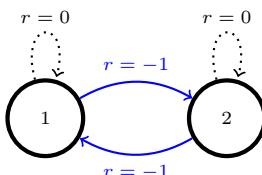
- (a) $\lim_{\gamma \rightarrow 1} P_\gamma^* = P^*$.
- (b) $\lim_{\gamma \rightarrow 1} \frac{P_\gamma^* - P^*}{1 - \gamma} = H$.

37.12 In this exercise you will prove the Part (a) of Theorem 37.1.

- (a) Prove there exists a deterministic stationary policy π^* and monotone increasing sequence of discount rates (γ_n) with $\gamma_n < 1$ and $\lim_{n \rightarrow \infty} \gamma_n = 1$ such that π^* is a greedy policy with respect to the fixed point v_n of T_{γ_n} for all n .
- (b) Show that $\rho^\pi = \rho\mathbf{1}$ is constant.
- (c) Let v be the value function such that $\rho\mathbf{1} + v = r_\pi + P_\pi v$. Show that (ρ, v) satisfies the Bellman optimality equation.

37.13 Consider the deterministic Markov decision process shown below with two states and two actions. The first action STAY keeps the state the same and the second action GO moves the learner to the other state while incurring a reward of negative one. Show that in this example solutions (ρ, v) to the Bellman optimality equations (Eq. (37.4)) are exactly the elements of the set

$$\{(\rho, v) \in \mathbb{R} \times \mathbb{R}^2 : \rho = 0, v(1) - 1 \leq v(2) \leq v(1) + 1\}.$$



37.14 Let M be a strongly connected MDP and (ρ^*, v) be a solution to the Bellman optimality equation. Show that $\text{span}(v) \leq (\rho^* - \min_{s,a} r_a(s))D(M)$.



Fix some states $s_1 \neq s_2$ and a memoryless policy π . Show that

$$v(s_2) - v(s_1) \leq (\rho^* - \min_{s,a} r_a(s))\mathbb{E}^\pi[\tau_{s_2} | S_1 = s_1].$$

Note for the sake of curiosity that the above display continues to hold for weakly communicating MDPs.



The proof of Theorem 4 in the paper by Bartlett and Tewari [2009] is incorrect, as is the sketch of the same result by Auer et al. [2010]. The problem is that the statement needs to hold for any solution v of the Bellman optimality equation. Both proofs use an argument that hinges on the fact that in an aperiodic

strongly connected MDP, v is in the set $\{c\mathbf{1} + \lim_{n \rightarrow \infty} T^n \mathbf{0} - n\rho^* : c \in \mathbb{R}\}$. However, Exercise 37.13 shows that there are some MDPs with the required properties where this does not hold.

37.15 Let M be a strongly connected MDP with rewards in $[0, 1]$, diameter $D < \infty$ and optimal gain ρ^* . Let $v_n^*(s)$ be the maximum total expected reward in n steps when the process starts in state s . Prove that $v_n^*(s) \leq n\rho^* + D$.

37.16 Let \tilde{M} be the extended MPD defined in Section 37.5.2. Prove that $P \in \mathcal{C}$ implies that \tilde{M} is strongly connected.

37.17 Prove that (37.9) follows from Theorem 37.4.

37.18 Fix state-space \mathcal{S} , action-space \mathcal{A} and reward function r . Let π be a policy with sublinear regret in all strongly connected MDPs $(\mathcal{S}, \mathcal{A}, r, P)$. Now suppose that $(\mathcal{S}, \mathcal{A}, r, P)$ is an MDP that is not strongly connected such that for all $s \in \mathcal{S}$ there exists state s' such is reachable from s under some policy and where $\rho_{s'}^* < \max_u \rho_u^*$. Finally, assume that $\rho_{S_1}^* = \max_u \rho_u^*$ almost surely. Prove that π has linear regret on this MDP.

37.19 This exercise develops the ideas mentioned in Note 10. First, we need some definitions: Fix \mathcal{S} and \mathcal{A} and define Π_0 as the set of policies (learner strategies) for MDPs with state space \mathcal{S} and action space \mathcal{A} that achieve sublinear regret in any strongly connected MDP with state space \mathcal{S} and action space \mathcal{A} . Now consider an arbitrary finite MDP $M = (\mathcal{S}, \mathcal{A}, P, r)$ that has \mathcal{S} as state space and \mathcal{A} as action space. A state $s \in \mathcal{S}$ is reachable from state $s' \in \mathcal{S}$ if there is a policy that, when started in s' reaches state s with positive probability after *one* or more steps. A set of states $C \subset \mathcal{S}$ is **strongly-connected component** (SCC) if every state $s \in C$ is reachable from every other state $s' \in C$ (allowing for the possibility that $s = s'$). Call C **maximal** if we cannot add more states to C and still maintain the SCC property. A maximal SCC is called a **maximal end-component** (MEC). Show the following:

- (a) Two MECs C_1 and C_2 are either equal, or disjoint.
- (b) Let C_1, \dots, C_k be all the distinct MECs of an MDP. The MDP structure defines a connectivity over C_1, \dots, C_k as follows: For $i \neq j$, we say that C_i is connected to C_j if from some state in C_i it is possible to reach some state of C_j with positive probability under some policy. Show that this connectivity structure defines a directed graph, which must be acyclic.
- (c) Let C_1, \dots, C_m with $m \leq k$ be the sinks (the nodes with no out-edges) of this graph. Show that if M is strongly connected then $m = 1$ and $C_1 = \mathcal{S}$.
- (d) Show that for any $i \in [m]$ and for any policy $\pi \in \Pi_0$ it holds that π will reach C_i in finite time with positive probability if the initial state distribution assigns positive mass to the non-trap states $\mathcal{S} \setminus \cup_{i \in [m]} C_i$.

- (e) Show that for $i \leq m$, for any $s \in C_i$ and any action $a \in \mathcal{A}$, $P_a(s, s') = 0$ for any $s' \in \mathcal{S} \setminus C_i$, i.e., C_i is **closed**.
(f) Show that the restriction of M to C_i defined as

$$M_i = (C_i, \mathcal{A}, (P_a(s))_{s \in C_i, a \in \mathcal{A}}, (r_a(s))_{s \in C_i, a \in \mathcal{A}})$$

is an MDP.

- (g) Show that M_i is strongly connected.
(h) Let τ be the time when the learner enters one of C_1, \dots, C_m and let $I \in [m]$ be the index of the class that is entered at time τ . That is, $S_\tau \in C_I$. Show that if M is strongly connected then $\tau = 1$ with probability one.
(i) We redefine the regret as follows: $R'_n = \mathbb{E} \left[\sum_{t=\tau}^{\tau+n-1} r_{A_t}(S_t) - n\rho^*(M_I) \right]$. Show that if M is strongly connected then $R_n = R'_n$.
(j) Show that there exist a learner such that (37.9) continues to hold in the sense that $R'_n \leq 1 + C\mathbb{E} \left[D(M_I)|C_I| \sqrt{2An \log(n)} \right]$.



The logic of the regret definition in Part (i) is that by Part (d), reasonable policies cannot control which trap they fall into in an MDP that has more than one traps. As such, policies should not be penalized for what trap they fall into. However, once a policy falls into some “trap”, we expect it to start to behave near optimally. What this definition is still lacking is that it is insensitive to how fast a policy gets trapped.

LEMMA 37.3 *Let (a_k) and (A_k) be nonnegative numbers so that for any $k \geq 0$, $a_{k+1} \leq A_k = 1 \vee (a_1 + \dots + a_k)$. Then for any $m \geq 1$,*

$$\sum_{k=1}^m \frac{a_k}{A_{k-1}} \leq (\sqrt{2} + 1) \sqrt{A_m}.$$

37.20 Prove Lemma 37.3.



Fix $(a_k)_k$ and $(A_k)_k$. Consider some $m \geq 1$. The statement is trivial if $\sum_{k=1}^{m-1} a_k \leq 1$. If this does not hold, use induction based on $m = n, n+1, \dots$ where n is the first integer such that $\sum_{k=1}^{n-1} a_k > 1$.

37.21 In this exercise you will modify the algorithm to handle the situation where r is unknown and rewards are stochastic. More precisely, assume there exists a function $r_a(s) \in [0, 1]$ for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$. Then in each round the learner observes S_t , chooses an action A_t and receives a reward $X_t \in [0, 1]$ with

$$\mathbb{E}[X_t | A_t, S_t] = r_{A_t}(S_t).$$

In order to accomodate the unknown reward function we modify UCRL2 in the

following way. First define the empirical reward at the start of the k th phase by

$$\hat{r}_{k,a}(s) = \sum_{u=1}^{\tau_k-1} \frac{\mathbb{I}\{S_u = s, A_u = a\} X_t}{1 \vee T_{\tau_k-1}(s, a)}.$$

Then let $\tilde{r}_{t,a}(s)$ be an upper confidence bound given by

$$\tilde{r}_{k,a}(s) = \hat{r}_{t,a}(s) + \sqrt{\frac{L}{2(1 \vee T_{\tau_k-1}(s, a))}},$$

where L is as in the proof of Theorem 37.4. The modified algorithm operates exactly like Algorithm 24, but replaces the unknown $r_a(s)$ with $\tilde{r}_{k,a}(s)$ when solving the extended MDP. Prove that with probability at least $1 - 3\delta/2$ the modified policy in the modified setting has regret at most

$$\hat{R}_n \leq CD(M)\mathcal{S} \sqrt{nA \log \left(\frac{nSA}{\delta} \right)},$$

where $C > 0$ is a universal constant.

37.22 The purpose of this exercise is to show that without phases UCRL2 may suffer linear regret. For convenience we consider the modified version of UCRL2 in Exercise 37.21 that does not know the reward. Now suppose we further modify this algorithm to re-solve the optimistic MDP in every round ($\tau_k = k$ for all k). We make use of the following deterministic Markov decision process with two actions $\mathcal{A} = \{\text{STAY}, \text{GO}\}$ represented by dashed and solid arrows respectively.

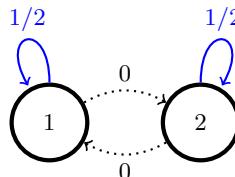


Figure 37.5 Transitions and rewards are deterministic. Numbers indicate the rewards.

- (a) Find all memoryless optimal policies for the MDP in Fig. 37.5.
- (b) Prove that the version of UCRL2 given in Exercise 37.21 modified to re-solve the optimistic MDP in every round suffers linear regret on this MDP.



Since UCRL2 and the environment are both deterministic you can examine the behavior of the algorithm on the MDP. You should aim to prove that eventually the algorithm will alternate between actions STAY and GO.

37.23 Design a simple algorithm for finding a state that is recurrent under some optimal policy using the solution to Eq. (37.5).



First solve Exercise 4.15 of the second volume of the book by [Bertsekas \[2012\]](#). Then adapt the argument and add some steps. Be warned this exercise is a little fiddly. The resulting algorithm should be much faster than solving all S versions of Eq. (37.6).

37.24 Consider a strongly connected MDP and suppose that ρ and v approximately satisfy the Bellman optimality equation in the sense that there exists an $\varepsilon > 0$ such that

$$\left| \rho + v(s) - \max_{a \in \mathcal{A}} r_a(s) - \langle P_a(s), v \rangle \right| \leq \varepsilon \quad \text{for all state/action pairs } s, a.$$

Let $\tilde{\pi}$ be the greedy policy with respect to v . Show that $\rho^{\tilde{\pi}}(s) \geq \bar{\rho}^\pi(s) - 2\varepsilon$ for all policies π .

37.25 In this exercise you will prove the claims to complete the proof of the lower bound.

- (a) Prove Claim 37.1.
- (b) Prove Claim 37.2.
- (c) Prove Claim 37.3.

37.26 Consider the MDP $M = (\mathcal{S}, \mathcal{A}, P, r)$ where $P_a(s) = p$ for some fixed categorical distribution p for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, where $\min_{s \in \mathcal{S}} p(s) > 0$. Assume that the rewards for action a in state s are sampled from a distribution supported on $[0, 1]$ (cf. note Item 4). An MDP like this defines nothing but a contextual bandit.

- (a) Derive the optimal policy and the average optimal reward.
- (b) Show an optimal value function that solves the Bellman optimality equation.
- (c) Prove that the diameter of this MDP is $D = \max_s 1/p(s)$.
- (d) Consider the algorithm that puts one instance of an appropriate version of UCB into every state (the same idea was explored in the context of adversarial bandits in Section 18.1). Prove that the expected regret of your algorithm will be at most $O(\sqrt{SAn})$.
- (e) Does the scaling behavior of the upper bound in Theorem 37.4 match the actual scaling behavior of the expected regret of UCRL2? Why or why not?
- (f) Design and run an experiment to confirm your claim.

37.27 This is a thinking and coding exercise to illustrate the difficulty of learning in Markov decision processes. The RiverSwim environment is originally due to [Strehl and Littman \[2008\]](#). The environment has two actions $\mathcal{A} = \{\text{LEFT}, \text{RIGHT}\}$ and $\mathcal{S} = [\text{S}]$ with $\text{S} \geq 2$. In all states $s > 1$, action LEFT deterministically leads to state $s - 1$ and provides no reward. In state 1, action LEFT leaves the state unchanged and yields a reward of 0.05. The action RIGHT tends to make the agent move right, but not deterministically (the learner is swimming against a

current). With probability 0.3 the state is incremented, with probability 0.6 the state is left unchanged, while with probability of 0.1 the state is decremented. This actions incurs reward zero in all states except in state S where it receives a reward of 1. The situation when S = 5 is illustrated in Fig. 37.6.

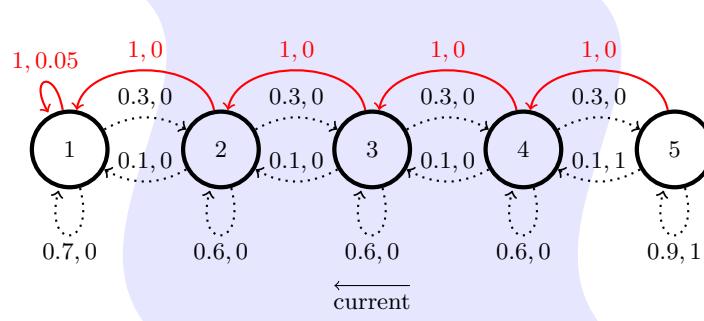


Figure 37.6 The RiverSwim MDP when S = 5. Solid arrows correspond to action LEFT and dashed ones to action RIGHT. The right-hand bank is slippery, so the learner sometimes falls back into the river.

- Show that the optimal policy always takes action RIGHT and calculate the optimal average reward ρ^* as a function of S.
- Implement the MDP and test the optimal policy when started from state 1. Plot the total reward as a function of time and compare it with the plot of $t \mapsto t\rho^*$. Run multiple simulations to produce error bars. How fast do you think the total reward concentrates around $t\rho^*$? Experiment with different values of S.
- The ε -greedy strategy can also be implemented in MDPs as follows: Based on the data previously collected estimate the transition probabilities and rewards using empirical means. Find the optimal policy π^* of the resulting MDP and if the current state is s , use the action $\pi^*(s)$ with probability $1 - \varepsilon$ and choose one of the two actions uniformly at random with the remaining probability. To ensure the empirical MDP has a well-defined optimal policy, mix the empirical estimate of the next state distributions $P_a(s)$ with the uniform distribution with a small mixture coefficient. Implement this strategy and plot the trajectories it exhibits for various MDP sizes. Explain what you see.
- Implement UCRL2 and produce the same plots. Can you explain what you see?
- Run simulations in RiverSwim instances of various sizes to compare the regret of UCRL2 and ε -greedy. What do you conclude?

This material will be published by Cambridge University Press as Bandit Algorithms by Tor Lattimore and Csaba Szepesvari. This pre-publication version is free to view and download for personal use only. Not for re-distribution, sale, or use in derivative works. © Tor Lattimore and Csaba Szepesvari 2017.

The latest version is available at <http://banditalgs.com>. Feedback on any aspect is very welcome: tor.lattimore@gmail.com

Appendix A Bibliography

- Y. Abbasi-Yadkori. *Forced-exploration based algorithms for playing in bandits with large action sets*. PhD thesis, University of Alberta, 2009a.
- Y. Abbasi-Yadkori. Forced-exploration based algorithms for playing in bandits with large action sets. Master's thesis, University of Alberta, Department of Computing Science, 2009b.
- Y. Abbasi-Yadkori. *Online Learning for Linearly Parametrized Control Problems*. PhD thesis, University of Alberta, 2012.
- Y. Abbasi-Yadkori and Cs. Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In S. M. Kakade and U. von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 1–26, Budapest, Hungary, 09–11 Jun 2011. PMLR.
- Y. Abbasi-Yadkori and Cs. Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, UAI, pages 2–11, Arlington, Virginia, United States, 2015. AUAI Press. ISBN 978-0-9966431-0-8.
- Y. Abbasi-Yadkori, A. Antos, and Cs. Szepesvári. Forced-exploration based algorithms for playing in stochastic linear bandits. In *COLT Workshop on On-line Learning with Limited Feedback*, 2009.
- Y. Abbasi-yadkori, D. Pál, and Cs. Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, NIPS, pages 2312–2320. Curran Associates, Inc., 2011.
- Y. Abbasi-Yadkori, D. Pal, and Cs. Szepesvári. Online-to-confidence-set conversions and application to sparse stochastic bandits. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1–9, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.
- Y. Abbasi-Yadkori, P. L. Bartlett, V. Kanade, Y. Seldin, and Cs. Szepesvári. Online learning in Markov decision processes with adversarially chosen transition probability distributions. In *Advances in Neural Information Processing Systems 26*, NIPS, pages 2508–2516, USA, 2013. Curran Associates Inc.

- N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the 16th International Conference on Machine Learning*, ICML, pages 3–11, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.
- M. Abeille and A. Lazaric. Linear Thompson sampling revisited. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 176–184, Fort Lauderdale, FL, USA, 20–22 Apr 2017a. PMLR.
- M. Abeille and A. Lazaric. Thompson sampling for linear-quadratic control problems. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1246–1254, Fort Lauderdale, FL, USA, 20–22 Apr 2017b. PMLR.
- J. D. Abernethy and A. Rakhlin. Beating the adaptive bandit with high probability. In *COLT*, 2009.
- J. D. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 263–274. Omnipress, 2008.
- J. D. Abernethy, E. Hazan, and A. Rakhlin. Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, 58(7):4164–4175, 2012.
- J. D. Abernethy, C. Lee, A. Sinha, and A. Tewari. Online linear optimization via smoothing. In M. F. Balcan, V. Feldman, and Cs. Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 807–823, Barcelona, Spain, 13–15 Jun 2014. PMLR.
- J. D. Abernethy, C. Lee, and A. Tewari. Fighting bandits with a new kind of smoothness. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, NIPS, pages 2197–2205. Curran Associates, Inc., 2015.
- M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964.
- L. Adelman. Choice theory. In Saul I. Gass and Michael C. Fu, editors, *Encyclopedia of Operations Research and Management Science*, pages 164–168. Springer US, Boston, MA, 2013.
- A. Agarwal, D. P. Foster, D. J. Hsu, S. M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, NIPS, pages 1035–1043. Curran Associates, Inc., 2011.
- A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on*

- Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1638–1646, Bejing, China, 22–24 Jun 2014. PMLR.
- Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, et al. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016.
- R. Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, pages 1054–1078, 1995.
- S. Agrawal and N. R. Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the 15th ACM conference on Economics and computation*, pages 989–1006. ACM, 2014.
- S. Agrawal and N. R. Devanur. Linear contextual bandits with knapsacks. In *Advances in Neural Information Processing Systems 29*, NIPS, pages 3458–3467. Curran Associates Inc., 2016.
- S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of Conference on Learning Theory (COLT)*, 2012.
- S. Agrawal and N. Goyal. Further optimal regret bounds for Thompson sampling. In C. M. Carvalho and P. Ravikumar, editors, *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 99–107, Scottsdale, Arizona, USA, 29 Apr–01 May 2013a. PMLR.
- S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 127–135, Atlanta, Georgia, USA, 17–19 Jun 2013b. PMLR.
- S. Agrawal and R. Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, NIPS, pages 1184–1194. Curran Associates, Inc., 2017.
- S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. Thompson sampling for the mnl-bandit. In S. Kale and O. Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 76–78, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.
- N. Ailon, Z. Karnin, and T. Joachims. Reducing dueling bandits to cardinal bandits. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, ICML’14, pages II–856–II–864. JMLR.org, 2014.
- J. Aldrich. “but you have to remember P. J. Daniell of Sheffield”. *Electronic Journal for History of Probability and Statistics*, 3(2), 2007.

- C. Allenberg, P. Auer, L. Györfi, and G. Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Proceedings of the 17th International Conference on Algorithmic Learning Theory*, ALT, pages 229–243, Berlin, Heidelberg, 2006. Springer-Verlag.
- N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the 28th annual ACM symposium on Theory of computing*, pages 20–29. ACM, 1996.
- N. Alon, N. Cesa-Bianchi, C. Gentile, and Y. Mansour. From bandits to experts: A tale of domination and independence. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, NIPS, pages 1610–1618. Curran Associates, Inc., 2013.
- N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren. Online learning with feedback graphs: Beyond bandits. In Peter GrÄijnwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 23–35, Paris, France, 03–06 Jul 2015. PMLR.
- V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, 1987.
- F. J. Anscombe. Sequential medical trials. *Journal of the American Statistical Association*, 58(302):365–383, 1963.
- A. Antos, G. Bartók, D. Pál, and Cs. Szepesvári. Toward a classification of finite partial-monitoring games. *Theoretical Computer Science*, 473:77–99, 2013.
- A. Arapostathis, V. S. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh, and S. I. Marcus. Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM Journal of Control and Optimization*, 31(2):282–344, 1993.
- Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. *arXiv preprint arXiv:1206.6400*, 2012.
- B. Ashwinkumar, J. Langford, and A. Slivkins. Resourceful contextual bandits. In M. F. Balcan, V. Feldman, and Cs. Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 1109–1134, Barcelona, Spain, 13–15 Jun 2014. PMLR.
- J.-V. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(Oct):2785–2836, 2010a.
- J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of Conference on Learning Theory (COLT)*, pages 217–226, 2009.
- J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *Proceedings of Conference on Learning Theory (COLT)*, 2010b.
- J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Tuning bandit algorithms in stochastic environments. In M. Hutter, R. A. Servedio, and E. Takimoto, editors, *Algorithmic Learning Theory*, pages 150–165, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

- J.-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.
- J.-Y. Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2013.
- P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.
- P. Auer and C. Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 116–120, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.
- P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 49–56. MIT Press, 2007.
- P. Auer and R. Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.
- P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 322–331. IEEE, 1995.
- P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002a.
- P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.
- P. Auer, R. Ortner, and Cs. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *International Conference on Computational Learning Theory*, pages 454–468. Springer, 2007.
- P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems 21*, NIPS, pages 89–96, 2009.
- P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, August 2010. ISSN 1532-4435.
- B. Awerbuch and R. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the 36th annual ACM symposium on Theory of computing*, pages 45–53. ACM, 2004.
- S. J. Axler. *Linear algebra done right*, volume 2. Springer, 1997.
- M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of*

- Machine Learning Research*, pages 263–272, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.
- A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 207–216. IEEE, 2013.
- K. Ball. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31:1–58, 1997.
- P. L. Bartlett and A. Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, UAI, pages 35–42, Arlington, Virginia, United States, 2009. AUAI Press.
- G. Bartók. A near-optimal algorithm for finite partial-monitoring games against adversarial opponents. In S. Shalev-Shwartz and I. Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 696–710. PMLR, 2013.
- G. Bartók, D. Pál, and Cs. Szepesvári. Toward a classification of finite partial-monitoring games. In *International Conference on Algorithmic Learning Theory*, pages 224–238. Springer, 2010.
- G. Bartók, N. Zolghadr, and Cs. Szepesvári. An adaptive algorithm for finite stochastic partial monitoring. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, ICML, pages 1779–1786, USA, 2012. Omnipress.
- G. Bartók, D. P. Foster, D. Pál, A. Rakhlin, and Cs. Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.
- J. A. Bather and H. Chernoff. Sequential decisions in the control of a spaceship. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 3, pages 181–207, 1967.
- R. Bellman. The theory of dynamic programming. Technical report, RAND CORP SANTA MONICA CA, 1954.
- R. E. Bellman. *Eye of the Hurricane*. World Scientific, 1984.
- D. Bernoulli. Exposition of a new theory on the measurement of risk. *Econometrica: Journal of the Econometric Society*, pages 23–36, 1954.
- A. C. Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.
- D. Berry and B. Fristedt. *Bandit problems : sequential allocation of experiments*. Chapman and Hall, London ; New York :, 1985.
- D. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, 1996.
- D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1-2. Athena Scientific, Belmont, MA, 4 edition, 2012.
- D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

- O. Besbes, Y. Gur, and A. Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 28*, NIPS, pages 199–207. Curran Associates, Inc., 2014.
- A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. E. Schapire. An optimal high probability algorithm for the contextual bandit problem. arXiv, 2010.
- A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire. Contextual bandit algorithms with supervised learning guarantees. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 19–26, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.
- P. Billingsley. *Probability and measure*. John Wiley & Sons, 2008.
- D. Blackwell. Controlled random walks. In *Proceedings of the international congress of mathematicians*, volume 3, pages 336–338, 1954.
- L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.
- S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.
- G. EP. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.
- G. EP. Box. Robustness in the strategy of scientific model building. *Robustness in statistics*, 1:201–236, 1979.
- S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- R. N. Bradt, S. M. Johnson, and S. Karlin. On sequential designs for maximizing the sum of n observations. *The Annals of Mathematical Statistics*, pages 1060–1074, 1956.
- R. Brafman and M. Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3: 213–231, 2003.
- J. Bretagnolle and C. Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.
- S. Bubeck and N. Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Foundations and Trends in Machine Learning. Now Publishers Incorporated, 2012.
- S. Bubeck and R. Eldan. The entropic barrier: a simple and optimal universal self-concordant barrier. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 279–279, Paris, France, 03–06 Jul 2015. PMLR.

- S. Bubeck and R. Eldan. Multi-scale exploration of convex functions and bandit convex optimization. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 583–589, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.
- S. Bubeck and C. Liu. Prior-free and prior-dependent regret bounds for Thompson sampling. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, NIPS, pages 638–646. Curran Associates, Inc., 2013.
- S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *COLT*, pages 42.1–42.23, 2012.
- S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.
- S. Bubeck, R. Munos, G. Stoltz, and Cs. Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, 2011.
- S. Bubeck, N. Cesa-Bianchi, and S. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Annual Conference on Learning Theory*, volume 23, pages 41–1. Microtome, 2012.
- S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *Information Theory, IEEE Transactions on*, 59(11):7711–7717, 2013a.
- S. Bubeck, V. Perchet, and P. Rigollet. Bounded regret in stochastic multi-armed bandits. In S. Shalev-Shwartz and I. Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 122–134, Princeton, NJ, USA, 12–14 Jun 2013b. PMLR.
- S. Bubeck, O. Dekel, T. Koren, and Y. Peres. Bandit convex optimization: \sqrt{T} regret in one dimension. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 266–278, Paris, France, 03–06 Jul 2015a. PMLR.
- S. Bubeck, R. Eldan, and J. Lehec. Finite-time analysis of projected langevin monte carlo. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, NIPS, pages 1243–1251. Curran Associates, Inc., 2015b.
- S. Bubeck, Y.T. Lee, and R. Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 72–85, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4528-6.
- S. Bubeck, M. Cohen, and Y. Li. Sparsity, variance and curvature in multi-armed bandits. In F. Janoos, M. Mohri, and K. Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 111–127. PMLR, 07–09 Apr 2018.

- A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.
- A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.
- A. N. Burnetas and M. N. Katehakis. Asymptotic Bayes analysis for the finite-horizon one-armed-bandit problem. *Probability in the Engineering and Informational Sciences*, 17(1):53–82, 2003.
- R. R. Bush and F. Mosteller. A stochastic model with applications to learning. *The Annals of Mathematical Statistics*, pages 559–585, 1953.
- O. Cappé, A. Garivier, O. Maillard, R. Munos, and G. Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.
- A. Carpentier and A. Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 590–604, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.
- A. Carpentier and R. Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 190–198, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.
- O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.
- N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.
- N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.
- N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31:562–580, 2006.
- N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora. Delay and cooperation in nonstochastic bandits. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 605–622, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.
- N. Cesa-Bianchi, C. Gentile, G. Lugosi, and G. Neu. Boltzmann exploration done right. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6284–6293. Curran Associates, Inc., 2017.
- J. Chakravorty and A. Mahajan. Multi-armed bandits, Gittins index, and its calculation. *Methods and Applications of Statistics in Clinical Trials: Planning, Analysis, and Inferential Methods, Volume 2*, pages 416–435, 2013.

- J. Chakravorty and A. Mahajan. Multi-armed bandits, Gittins index, and its calculation. *Methods and Applications of Statistics in Clinical Trials: Planning, Analysis, and Inferential Methods, Volume 2*, pages 416–435, 2014.
- J. T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997.
- O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, NIPS, pages 2249–2257. Curran Associates, Inc., 2011.
- Y. R. Chen and M. N. Katehakis. Linear programming for finite state multi-armed bandit problems. *Mathematics of Operations Research*, 11(1):180–183, 1986.
- H. Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.
- H. Chernoff. A career in statistics. *Past, Present, and Future of Statistical Science*, page 29, 2014.
- W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.
- A. Chuklin, I. Markov, and M. de Rijke. *Click Models for Web Search*. Morgan & Claypool Publishers, 2015.
- A. Cohen and T. Hazan. Following the perturbed leader for online structured learning. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1034–1042, Lille, France, 07–09 Jul 2015. PMLR.
- A. Cohen, T. Hazan, and T. Koren. Tight bounds for bandit combinatorial optimization. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 629–642, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.
- R. Combes, S. Magureanu, A. Proutiere, and C. Laroche. Learning to rank: Regret lower bounds and efficient algorithms. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 231–244. ACM, 2015. ISBN 978-1-4503-3486-0.
- R. Combesd, M. Shahi, A. Proutiere, and M. Lelarge. Combinatorial bandits revisited. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, NIPS, pages 2116–2124. Curran Associates, Inc., 2015.
- T. M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, 1991.
- T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- W. Cowan and M. N. Katehakis. An asymptotically optimal policy for uniform bandits of unknown support. *arXiv preprint arXiv:1505.01918*, 2015.

- W. Cowan, J. Honda, and M. N. Katehakis. Normal bandits of unknown means and variances: Asymptotic optimality, finite horizon regret bounds, and a solution to an open problem. *arXiv preprint arXiv:1504.05823*, 2015.
- K. Crammer and C. Gentile. Multiclass classification with bandit feedback using adaptive regularization. *Machine learning*, 90(3):347–383, 2013.
- N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 87–94. ACM, 2008.
- V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of Conference on Learning Theory*, COLT, pages 355–366, 2008.
- R. Degenne and V. Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1587–1595, New York, New York, USA, 20–22 Jun 2016. PMLR.
- O. Dekel, C. Gentile, and K. Sridharan. Robust selective sampling from single and multiple teachers. In *COLT*, pages 346–358, 2010.
- O. Dekel, C. Gentile, and K. Sridharan. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, 13 (Sep):2655–2697, 2012.
- A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38. Springer Science & Business Media, 2009.
- E. V. Denardo, H. Park, and U. G. Rothblum. Risk-Sensitive and Risk-Neutral Multiarmed Bandits. *Mathematics of Operations Research*, 32(2):374–394, 2007.
- T. Desautels, A. Krause, and J. W. Burdick. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *Journal of Machine Learning Research*, 15:4053–4103, 2014.
- S. Dong and B. Van Roy. An information-theoretic analysis for thompson sampling with many actions. *arXiv preprint arXiv:1805.11845*, 2018.
- J. L. Doob. *Stochastic processes*. Wiley, 1953.
- M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, UAI, pages 169–178. AUAI Press, 2011.
- M. Dudík, K. Hofmann, R. E. Schapire, A. Slivkins, and M. Zoghi. Contextual dueling bandits. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 563–587, Paris, France, 03–06 Jul 2015. PMLR.
- R. M. Dudley. *Uniform central limit theorems*, volume 142. Cambridge university press, 2014.
- C. G. Esseen. *On the Liapounoff limit of error in the theory of probability*. Almqvist & Wiksell, 1942.

- E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *Computational Learning Theory*, pages 255–270. Springer, 2002.
- E. Even-Dar, S. M. Kakade, and Y. Mansour. Experts in a Markov decision process. In *Advances in Neural Information Processing Systems 17*, NIPS, pages 401–408, Cambridge, MA, USA, 2004. MIT Press.
- E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(Jun):1079–1105, 2006.
- S. Filippi, O. Cappe, A. Garivier, and Cs. Szepesvári. Parametric bandits: The generalized linear case. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, NIPS, pages 586–594. Curran Associates, Inc., 2010.
- D. Foster and A. Rakhlin. No internal regret via neighborhood watch. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 382–390, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.
- S. Frederick, G. Loewenstein, and T. O’donoghue. Time discounting and time preference: A critical review. *Journal of economic literature*, 40(2):351–401, 2002.
- E. Frostig and G. Weiss. Four proofs of Gittins’ multiarmed bandit theorem. *Applied Probability Trust*, 70, 1999.
- Y. Gai, B. Krishnamachari, and R. Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, 2012.
- A. Garivier. Informational confidence bounds for self-normalized averages and applications. *arXiv preprint arXiv:1309.3376*, 2013.
- A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of Conference on Learning Theory (COLT)*, 2011.
- A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.
- A. Garivier and E. Moulines. On upper-confidence bound policies for switching bandit problems. In J. Kivinen, Cs. Szepesvári, E. Ukkonen, and T. Zeugmann, editors, *Algorithmic Learning Theory*, pages 174–188, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.
- A. Garivier, E. Kaufmann, and W. M. Koolen. Maximin action identification: A new bandit framework for games. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1028–1050, Columbia University, New York, New York, USA, 23–26 Jun 2016a. PMLR.

- A. Garivier, T. Lattimore, and E. Kaufmann. On explore-then-commit strategies. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, NIPS, pages 784–792. Curran Associates, Inc., 2016b.
- A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *arXiv preprint arXiv:1602.07182*, 2016c.
- A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.
- C. Gentile and F. Orabona. On multilabel classification and ranking with partial feedback. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, NIPS, pages 1151–1159. Curran Associates, Inc., 2012.
- C. Gentile and F. Orabona. On multilabel classification and ranking with bandit feedback. *Journal of Machine Learning Research*, 15(1):2451–2487, 2014.
- S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14(Mar):729–769, 2013.
- S. Gerchinovitz and T. Lattimore. Refined lower bounds for adversarial bandits. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, NIPS, pages 1198–1206. Curran Associates, Inc., 2016.
- S. Ghosal and A. van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.
- J. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.
- J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.
- A. Gopalan and S. Mannor. Thompson sampling for learning parameterized Markov decision processes. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 861–898, Paris, France, 03–06 Jul 2015. PMLR.
- T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. Web-scale Bayesian click-through rate prediction for sponsored search advertising in microsoft’s bing search engine. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML, pages 13–20, USA, 2010. Omnipress.
- O. Granmo. Solving two-armed bernoulli bandit problems using a Bayesian learning automaton. *International Journal of Intelligent Computing and Cybernetics*, 3(2):207–234, 2010.
- R. M. Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.
- K. Greenewald, A. Tewari, S. Murphy, and P. Klasnja. Action centered contextual bandits. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus,

- S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5977–5985. Curran Associates, Inc., 2017.
- M. Grötschel, L. Lovász, and A. Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.
- F. Guo, C. Liu, and Y. M. Wang. Efficient multiple-click models in web search. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, pages 124–131. ACM, 2009.
- L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, 2002.
- A. György and Cs. Szepesvári. Shifting regret, mirror descent, and matrices. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2943–2951, New York, New York, USA, 20–22 Jun 2016. PMLR.
- A. György, T. Linder, G. Lugosi, and G. Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8 (Oct):2369–2403, 2007.
- M. Hanawal, V. Saligrama, M. Valko, and R. Munos. Cheap bandits. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2133–2142, Lille, France, 07–09 Jul 2015. PMLR.
- J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.
- Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.
- E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.
- E. Hazan and S. Kale. A simple multi-armed bandit algorithm with optimal variation-bounded regret. In S. M. Kakade and U. von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 817–820. PMLR, 2011.
- E. Hazan, Z. Karnin, and R. Meka. Volumetric spanners: an efficient exploration basis for learning. *Journal of Machine Learning Research*, 17(119):1–34, 2016.
- M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.
- M. Herbster and M. K. Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1(Sep):281–309, 2001.
- J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Proceedings of Conference on Learning Theory (COLT)*, pages 67–79, 2010.
- J. Honda and A. Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361–391, 2011.
- J. Honda and A. Takemura. Optimality of Thompson sampling for Gaussian bandits depends on priors. In S. Kaski and J. Corander, editors, *Proceedings of*

- the 17th International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 375–383, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.
- J. Honda and A. Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Journal of Machine Learning Research*, 16:3721–3756, 2015.
- R. Huang, M. M. Ajallooeian, Cs. Szepesvári, and M. Müller. Structured best arm identification with fixed confidence. In S. Hanneke and L. Reyzin, editors, *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76 of *Proceedings of Machine Learning Research*, pages 593–616, Kyoto University, Kyoto, Japan, 2017a. PMLR.
- R. Huang, T. Lattimore, A. György, and Cs. Szepesvári. Following the leader and fast rates in online linear prediction: Curved constraint sets and other regularities. *Journal of Machine Learning Research*, 18:1–31, 2017b.
- M. Hutter and J. Poland. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research*, 6:639–660, 2005.
- E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.
- K. Jamieson and R. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *Information Sciences and Systems (CISS), 2014 48th Annual Conference on*, pages 1–6. IEEE, 2014.
- K. Jamieson, S. Katoriya, A. Deshpande, and R. Nowak. Sparse dueling bandits. In G. Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 416–424, San Diego, California, USA, 09–12 May 2015. PMLR.
- E. T. Jaynes. *Probability theory: the logic of science*. Cambridge university press, 2003.
- F. John. Extremum problems with inequalities as subsidiary conditions. *Courant Anniversary Volume, Interscience*, 1948.
- P. Joulani, A. Gyorgy, and Cs. Szepesvari. Online learning under delayed feedback. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1453–1461, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.
- K. Jun, A. Bhargava, R. Nowak, and R. Willett. Scalable generalized linear bandits: Online computation and hashing. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 99–109. Curran Associates, Inc., 2017.
- L. P. Kaelbling. *Learning in embedded systems*. MIT press, 1993.
- W. Kahan. Pracniques: further remarks on reducing truncation errors. *Communications of the ACM*, 8(1):40, 1965.

- D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–91, 1979.
- S. Kakade. *On The Sample Complexity Of Reinforcement Learning*. PhD thesis, University College London, 2003.
- S. M. Kakade, S. Shalev-Shwartz, and A. Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th International Conference on Machine Learning*, pages 440–447, 2008.
- A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005a.
- A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005b.
- L. Kallenberg. A note on M.N. Katehakis' and Y.-R. Chen's computation of the Gittins index. *Mathematics of operations research*, 11(1):184–186, 1986.
- L. Kallenberg. Markov decision processes: Lecture notes. 2016.
- O. Kallenberg. *Foundations of modern probability*. Springer-Verlag, 2002.
- Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1238–1246, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.
- S. Katariya, B. Kveton, Cs. Szepesvári, and Z. Wen. DCM bandits: Learning to rank with multiple clicks. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1215–1224, 2016.
- S. Katariya, B. Kveton, Cs. Szepesvári, C. Vernade, and Z. Wen. Bernoulli rank-1 bandits for click feedback. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017a.
- S. Katariya, B. Kveton, Cs. Szepesvári, C. Vernade, and Z. Wen. Stochastic rank-1 bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017b.
- M. N. Katehakis and H. Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92(19):8584, 1995.
- E. Kaufmann. On Bayesian index policies for sequential resource allocation. *The Annals of Statistics*, 46(2):842–865, 04 2018.
- E. Kaufmann, O. Cappe, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 592–600, La Palma, Canary Islands, 21–23 Apr 2012a. PMLR.
- E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In NaderH. Bshouty, Gilles Stoltz, Nicolas Vayatis, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, pages 199–213. Springer Berlin Heidelberg, 2012b. ISBN 978-3-642-34105-2.

- J. Kawale, H. H. Bui, B. Kveton, L. Tran-Thanh, and S. Chawla. Efficient Thompson sampling for online matrix-factorization recommendation. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, NIPS, pages 1297–1305. Curran Associates, Inc., 2015.
- A. Kazerouni, M. Ghavamzadeh, Y. Abbasi, and B. Van Roy. Conservative contextual linear bandits. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3910–3919. Curran Associates, Inc., 2017.
- M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.
- M. J. Kearns and U. V. Vazirani. *An introduction to computational learning theory*. MIT press, 1994.
- J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12(5):363–365, 1960.
- M. J. Kim. Thompson sampling for stochastic control: The finite parameter case. *IEEE Transactions on Automatic Control*, 62(12):6415–6422, 2017.
- J. Kirschner and A. Krause. Information directed sampling and bandits with heteroscedastic noise. *arXiv preprint arXiv:1801.09667*, 2018.
- R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, NIPS, pages 697–704. MIT Press, 2005.
- R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 681–690. ACM, 2008.
- T. Kocák, G. Neu, M. Valko, and R. Munos. Efficient learning by implicit exploration in bandit problems with side observations. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 613–621. Curran Associates, Inc., 2014.
- T. Kocák, M. Valko, R. Munos, and S. Agrawal. Spectral Thompson sampling. In *AAAI*, pages 1911–1917, 2014.
- L. Kocsis and Cs. Szepesvári. Discounted UCB. In *2nd PASCAL Challenges Workshop*, pages 784–791, 2006.
- J. Komiyama, J. Honda, H. Kashima, and H. Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1141–1154, Paris, France, 03–06 Jul 2015a. PMLR.
- J. Komiyama, J. Honda, and H. Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural*

- Information Processing Systems 28*, NIPS, pages 1792–1800. Curran Associates, Inc., 2015b.
- W. M. Koolen, M. K. Warmuth, and J. Kivinen. Hedging structured concepts. In *COLT*, pages 93–105. Omnipress, 2010.
- N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for 1-dimensional exponential family bandits. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1448–1456. Curran Associates, Inc., 2013.
- S. R. Kulkarni and G. Lugosi. Finite-time lower bounds for the two-armed bandit problem. *IEEE Transactions on Automatic Control*, 45(4):711–714, 2000.
- B. Kveton, Cs. Szepesvári, Z. Wen, and A. Ashkan. Cascading bandits: Learning to rank in the cascade model. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, pages 767–776. JMLR.org, 2015a.
- B. Kveton, Z. Wen, A. Ashkan, and Cs. Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In G. Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 535–543, San Diego, California, USA, 09–12 May 2015b. PMLR.
- B. Kveton, Z. Wen, Z. Ashkan, and Cs. Szepesvári. Combinatorial cascading bandits. In *Advances in Neural Information Processing Systems 28*, NIPS, pages 1450–1458. Curran Associates Inc., 2015c.
- P. Lagree, C. Vernade, and O. Cappé. Multiple-play bandits in the position-based model. In *Advances in Neural Information Processing Systems 29*, NIPS, pages 1597–1605. Curran Associates Inc., 2016.
- T. L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987.
- T. L. Lai. Martingales in sequential analysis and time series, 1945–1985. *Electronic Journal for history of probability and statistics*, 5(1), 2009.
- T. L. Lai and T. Graves. Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743, 1997.
- T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.
- J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, NIPS, pages 817–824. Curran Associates, Inc., 2008.
- P. Laplace. *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator*, volume 13. Springer Science & Business Media, 2012.
- T. Lattimore. The pareto regret frontier for bandits. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in*

- Neural Information Processing Systems 28*, NIPS, pages 208–216. Curran Associates, Inc., 2015a.
- T. Lattimore. Optimally confident UCB: Improved regret for finite-armed bandits. *arXiv preprint arXiv:1507.07880*, 2015b.
- T. Lattimore. Regret analysis of the finite-horizon Gittins index strategy for multi-armed bandits. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1214–1245, Columbia University, New York, New York, USA, 23–26 Jun 2016a. PMLR.
- T. Lattimore. Regret analysis of the anytime optimally confident UCB algorithm. Technical report, University of Alberta, 2016b.
- T. Lattimore. A scale free algorithm for stochastic bandits with bounded kurtosis. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1584–1593. Curran Associates, Inc., 2017.
- T. Lattimore. Refining the confidence level for optimistic bandit strategies. *technical report*, 2018.
- T. Lattimore and M. Hutter. PAC bounds for discounted MDPs. In Nicolas Vayatis Nader H. Bshouty, Gilles Stoltz and Thomas Zeugmann, editors, *Proceedings of the 23th International Conference on Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, pages 320–334. Springer Berlin / Heidelberg, 2012.
- T. Lattimore and R. Munos. Bounded regret for finite-armed structured bandits. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, NIPS, pages 550–558. Curran Associates, Inc., 2014.
- T. Lattimore and Cs. Szepesvári. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 728–737, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.
- T. Lattimore and Cs. Szepesvári. Cleaning up the neighbourhood: A full classification for adversarial partial monitoring. *submitted*, 2018.
- T. Lattimore, K. Crammer, and Cs. Szepesvári. Linear multi-resource allocation with semi-bandit feedback. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, NIPS, pages 964–972. Curran Associates, Inc., 2015.
- T. Lattimore, B. Kveton, S. Li, and Cs. Szepesvári. Toprank: A practical algorithm for online stochastic ranking. *submitted*, 2018.
- B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.
- A. Lazaric and R. Munos. Hybrid stochastic-adversarial on-line learning. In *COLT*, 2009.

- T. Le, Cs. Szepesvari, and R. Zheng. Sequential learning for multi-channel wireless network monitoring with channel switching costs. *IEEE Transactions on Signal Processing*, 62(22):5919–5929, 2014.
- E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.
- H. Lei, A. Tewari, and S. A. Murphy. An actor-critic contextual bandit algorithm for personalized mobile health interventions. 2017.
- J. Leike, T. Lattimore, L. Orseau, and M. Hutter. Thompson sampling is asymptotically optimal in general environments. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, UAI, pages 417–426. AUAI Press, 2016.
- H. R. Lerche. *Boundary crossing of Brownian motion: Its relation to the law of the iterated logarithm and to sequential analysis*. Springer, 1986.
- D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.
- S. Li, B. Wang, S. Zhang, and W. Chen. Contextual combinatorial cascading bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1245–1253, 2016.
- N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.
- L. Lovasz and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.
- D. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
- O. Maillard. Robust risk-averse stochastic multi-armed bandits. In *ALT*, pages 218–233. Springer, Berlin, Heidelberg, 2013.
- O. Maillard, R. Munos, and G. Stoltz. Finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of Conference On Learning Theory (COLT)*, 2011.
- S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 684–692. Curran Associates, Inc., 2011.
- S. Mannor and N. Shimkin. On-line learning with imperfect monitoring. In *Learning Theory and Kernel Machines*, pages 552–566. Springer, 2003.
- S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, December 2004.
- S. Mannor, V. Perchet, and G. Stoltz. Set-valued approachability and online learning with partial monitoring. *The Journal of Machine Learning Research*, 15(1):3247–3295, 2014.
- H. Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.
- M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7(3):216–244, 1960.

- A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.
- B. C. May, N. Korda, A. Lee, and D. S. Leslie. Optimistic Bayesian sampling in contextual-bandit problems. *The Journal of Machine Learning Research*, 13(1):2069–2106, 2012.
- C. McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.
- H. B. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *COLT*, volume 3120, pages 109–123. Springer, 2004.
- H. B. McMahan and M. J. Streeter. Tighter bounds for multi-armed bandits with expert advice. In *COLT*, 2009.
- P. Ménard and A. Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In S. Hanneke and L. Reyzin, editors, *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76 of *Proceedings of Machine Learning Research*, pages 223–237, Kyoto University, Kyoto, Japan, 15–17 Oct 2017. PMLR.
- S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.
- V. Mnih, Cs. Szepesvári, and J.-Y. Audibert. Empirical bernstein stopping. In *Proceedings of the 25th International Conference on Machine Learning*, ICML, pages 672–679, New York, NY, USA, 2008. ACM.
- M. I. Muller, P. E. Valenzuela, A. Proutiere, and C. R. Rojas. A stochastic multi-armed bandit approach to nonparametric h_∞ -norm estimation. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 4632–4637. IEEE, 2017.
- R. Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, NIPS, pages 783–791. Curran Associates, Inc., 2011.
- J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.
- A. Nemirovski. Efficient methods for large-scale convex optimization problems. *Ekonomika i Matematicheskie Metody*, 15, 1979.
- A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.
- G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, NIPS, pages 3168–3176. Curran Associates, Inc., 2015a.
- G. Neu. First-order regret bounds for combinatorial semi-bandits. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1360–1375, Paris, France, 03–06 Jul 2015b. PMLR.

- G. Neu, A. György, Cs. Szepesvári, and A. Antos. Online Markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59(3):676–691, December 2014.
- J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1944.
- J. Niño-Mora. Computing a classic index for finite-horizon bandits. *INFORMS Journal on Computing*, 23(2):254–267, 2011.
- P. A. Ortega and D. A. Braun. A minimum relative entropy principle for learning and acting. *Journal of Artificial Intelligence Research*, 38:475–511, 2010.
- R. Ortner and D. Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *Advances in Neural Information Processing Systems 25*, NIPS, pages 1763–1771, USA, 2012. Curran Associates Inc.
- R. Ortner, D. Ryabko, P. Auer, and R. Munos. Regret bounds for restless Markov bandits. In N. Bshouty, G. Stoltz, N. Vayatis, and T. Zeugmann, editors, *Algorithmic Learning Theory*, pages 214–228, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.
- I. Osband and B. Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2701–2710, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.
- I. Osband, D. Russo, and B. Van Roy. (more) efficient reinforcement learning via posterior sampling. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, NIPS, pages 3003–3011. Curran Associates, Inc., 2013.
- E. Ostrovsky and L. Sirota. Exact value for subgaussian norm of centered indicator random variable. *arXiv preprint arXiv:1405.6749*, 2014.
- C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.
- V. H. Peña, T.L. Lai, and Q. Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.
- V. Perchet. Approachability of convex sets in games with partial monitoring. *Journal of Optimization Theory and Applications*, 149(3):665–677, 2011.
- G. Peskir and A. Shiryaev. *Optimal stopping and free-boundary problems*. Springer, 2006.
- A. Piccolboni and C. Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *Computational Learning Theory*, pages 208–223. Springer, 2001.
- Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4102–4110, StockholmsmÅdssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

- J. Poland. FPL analysis for adaptive bandits. In O. B. Luponov, O. M. Kasim-Zade, A. V. Chaskin, and K. Steinhöfel, editors, *Stochastic Algorithms: Foundations and Applications*, pages 58–69, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.
- D. Pollard. *A user's guide to measure theoretic probability*, volume 8. Cambridge University Press, 2002.
- M. Puterman. *Markov decision processes: discrete stochastic dynamic programming*, volume 414. Wiley, 2009.
- F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning*, pages 784–791. ACM, 2008.
- A. N. Rafferty, H. Ying, and J. J. Williams. Bandit assignment for educational experiments: Benefits to students versus statistical power. In *Artificial Intelligence in Education*, pages 286–290. Springer, 2018.
- A. Rakhlin and K. Sridharan. BISTRO: An efficient relaxation-based method for contextual bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1977–1985, 2016.
- A. Rakhlin and K. Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. In S. Kale and O. Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1704–1722, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.
- H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.
- H. Robbins and D. Siegmund. Boundary crossing probabilities for the wiener process and sample sums. *The Annals of Mathematical Statistics*, pages 1410–1429, 1970.
- H. Robbins, D. Sigmund, and Y. Chow. Great expectations: the theory of optimal stopping. *Houghton-Niffin*, 7:631–640, 1971.
- S. Robertson. The probability ranking principle in IR. 33:294–304, 12 1977.
- R. T. Rockafellar. *Convex analysis*. Princeton university press, 2015.
- R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.
- C. A. Rogers. *Packing and covering*. Cambridge University Press, 1964.
- S. M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, New York, 1983.
- P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.
- D. Russo. Simple Bayesian algorithms for best arm identification. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1417–1418, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

- D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, NIPS, pages 2256–2264. Curran Associates, Inc., 2013.
- D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, NIPS, pages 1583–1591. Curran Associates, Inc., 2014a.
- D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014b.
- D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(1):2442–2471, 2016. ISSN 1532-4435.
- D. Russo, B. Van Roy, A. Kazerouni, and I. Osband. A tutorial on Thompson sampling. *arXiv preprint arXiv:1707.02038*, 2017.
- A. Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29(1):224–243, 1999.
- A. Salomon, J. Audibert, and I. Alaoui. Lower bounds and selectivity of weak-consistent policies in stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 14(Jan):187–207, 2013.
- P. Samuelson. A note on measurement of utility. *The Review of Economic Studies*, 4(2):pp. 155–161, 1937.
- A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3275–3283. Curran Associates, Inc., 2012.
- Y. Seldin and G. Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In S. Kale and O. Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1743–1759, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.
- Y. Seldin and A. Slivkins. One practical algorithm for both stochastic and adversarial bandits. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1287–1295, Bejing, China, 22–24 Jun 2014. PMLR.
- S. Shalev-Shwartz. *Online learning: Theory, algorithms, and applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.
- S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2009.
- S. Shalev-Shwartz and Y. Singer. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69(2-3):115–142, 2007.
- O. Shamir. On the complexity of bandit linear optimization. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning*

- Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1523–1551, Paris, France, 03–06 Jul 2015. PMLR.
- D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, and M. Lanctot. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
- S. D. Silvey and B. Sibson. Discussion of dr. wynn’s and of dr. laycock’s papers. *Journal of Royal Statistical Society (B)*, 34:174–175, 1972.
- M. Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.
- A. Slivkins. Contextual bandits with similarity information. *Journal of Machine Learning Research*, 15(1):2533–2568, 2014.
- A. Slivkins and E. Upfal. Adapting to a changing environment: the brownian restless bandits. In *COLT*, pages 343–354, 2008.
- M. Soare, A. Lazaric, and R. Munos. Best-arm identification in linear bandits. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, NIPS, pages 828–836. Curran Associates, Inc., 2014.
- I. M. Sonin. A generalized Gittins index for a Markov chain and its recursive calculation. *Statistics & Probability Letters*, 78(12):1526–1533, 2008.
- N. Srebro, K. Sridharan, and A. Tewari. On the universality of online mirror descent. In *Advances in neural information processing systems*, pages 2645–2653, 2011.
- K. Sridharan and A. Tewari. Convex games in banach spaces. In *Proceedings of the 23rd Conference on Learning Theory*, pages 1–13. Omnipress, 2010.
- G. Stoltz. *Incomplete information and internal regret in prediction of individual sequences*. PhD thesis, Université Paris Sud-Paris XI, 2005.
- R. E. Strauch. Negative dynamic programming. *The Annals of Mathematical Statistics*, 37(4):871–890, 08 1966.
- A. Strehl and M. Littman. A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd international conference on Machine learning*, ICML, pages 856–863, New York, NY, USA, 2005. ACM.
- A. Strehl and M. Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.
- A. Strehl, L. Li, E. Wiewiora, J. Langford, and M. Littman. PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888, New York, NY, USA, 2006. ACM.
- M. J. A. Strens. A Bayesian framework for reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, ICML ’00, pages 943–950, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-707-2.
- Y. Sui, A. Gotovos, J. Burdick, and A. Krause. Safe exploration for optimization with gaussian processes. In Francis Bach and David Blei, editors, *Proceedings*

- of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 997–1005, Lille, France, 07–09 Jul 2015. PMLR.
- Q. Sun, W. Zhou, and J. Fan. Adaptive huber regression: Optimality and phase transition. *arXiv preprint arXiv:1706.06991*, 2017.
- R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.
- V. Syrgkanis, A. Krishnamurthy, and R. Schapire. Efficient algorithms for adversarial contextual learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2159–2168, New York, New York, USA, 2016. PMLR.
- Cs. Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.
- I. Szita and A. Lörincz. Optimistic initialization and greediness lead to polynomial time learning in factored MDPs. In *Proceedings of the 26th International Conference on Machine Learning*, ICML ’09, pages 1001–1008, New York, NY, USA, 2009. ACM.
- I. Szita and Cs. Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML’10, pages 1031–1038, USA, 2010. Omnipress. ISBN 978-1-60558-907-7.
- M. Talagrand. The missing factor in Hoeffding’s inequalities. *Annales de l’IHP Probabilités et statistiques*, 31(4):689–702, 1995.
- J. Teevan, S. T. Dumais, and E. Horvitz. Characterizing the value of personalizing search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 757–758, New York, NY, USA, 2007. ACM.
- A. Tewari and P. L. Bartlett. Optimistic linear programming gives logarithmic regret for irreducible mdps. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1505–1512. Curran Associates, Inc., 2008.
- A. Tewari and S. A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.
- G. Theocharous, Z. Wen, Y. Abbasi-Yadkori, and N. Vlassis. Posterior sampling for large scale reinforcement learning. *arXiv preprint arXiv:1711.07979*, 2017.
- W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.
- W. R. Thompson. On the theory of apportionment. *American Journal of Mathematics*, 57(2):450–456, 1935.
- M. J. Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.
- J. R. R. Tolkien. *The Hobbit*. Ballantine Books, 1937.
- L. Tran-Thanh, A. Chapman, E. Munoz de Cote, A. Rogers, and N. R. Jennings. Epsilon- $\tilde{\Delta}$ first policies for budget- $\tilde{\Delta}$ limited multi-armed bandits.

- In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, AAAI, pages 1211–1216, 2010.
- L. Tran-Thanh, A. Chapman, A. Rogers, and N. R. Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, AAAI’12, pages 1134–1140. AAAI Press, 2012.
- J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.
- J. N. Tsitsiklis. A short proof of the Gittins index theorem. *The Annals of Applied Probability*, pages 194–199, 1994.
- A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.
- C. Ionescu Tulcea. Mesures dans les espaces produits. *Atti Accad. Naz. Lincei Rend.*, 7:208–211, 1949–50.
- E. Uchibe and K. Doya. Competitive-cooperative-concurrent reinforcement learning with importance sampling. In *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals and Animats*, pages 287–296, 2004.
- A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996.
- M. Valko. Bandits on graphs and structures, 2016.
- M. Valko, A. Carpentier, and R. Munos. Stochastic simultaneous optimistic optimization. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 19–27, Atlanta, Georgia, USA, 17–19 Jun 2013a. PMLR.
- M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, UAI, pages 654–663, Arlington, Virginia, United States, 2013b. AUAI Press.
- M. Valko, R. Munos, B. Kveton, and T. Kocák. Spectral bandits for smooth graph functions. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 46–54, Bejing, China, 22–24 Jun 2014. PMLR.
- S. van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.
- D. van der Hoeven, T. van Erven, and W. Kotłowski. The many faces of exponential weights in online learning. *arXiv preprint arXiv:1802.07543*, 2018.
- H. P. Vanchinathan, G. Bartók, and A. Krause. Efficient partial monitoring with prior information. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, NIPS, pages 1691–1699. Curran Associates, Inc., 2014.

- L. Vandenberghe, S. Boyd, and S.-P. Wu. Determinant maximization with linear matrix inequality constraints. *SIAM journal on matrix analysis and applications*, 19(2):499–533, 1998.
- P. Varaiya, J. Walrand, and C. Buyukkoc. Extensions of the multiarmed bandit problem: The discounted case. *IEEE transactions on automatic control*, 30(5):426–439, 1985.
- C. Vernade, O. Cappé, and V. Perchet. Stochastic bandit models for delayed conversions. *arXiv preprint arXiv:1706.09186*, 2017.
- C. Vernade, A. Carpentier, G. Zappella, B. Ermis, and M. Brueckner. Contextual bandits under delayed feedback. *arXiv preprint arXiv:1807.02089*, 2018.
- S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199–215, 2015.
- W. Vogel. An asymptotic minimax theorem for the two armed bandit problem. *The Annals of Mathematical Statistics*, 31(2):444–451, 1960.
- J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.
- V. G. Vovk. Aggregating strategies. *Proceedings of Computational Learning Theory*, 1990.
- P. L Wawrzynski and A. Pacut. Truncated importance sampling for reinforcement learning with experience replay. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, pages 305–315, 2007.
- R. Weber. On the Gittins index for multiarmed bandits. *The Annals of Applied Probability*, 2(4):1024–1033, 1992.
- R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.
- M. J. Weinberger and E. Ordentlich. On delayed prediction of individual sequences. In *Information Theory, 2002. Proceedings. 2002 IEEE International Symposium on*, page 148. IEEE, 2002.
- T. Weissman, E. Ordentlich, G. Seroussi, and S. Verdú. Inequalities for the ℓ^1 deviation of the empirical distribution. Technical report, Hewlett-Packard Labs, 2003.
- Z. Wen, B. Kveton, and A. Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1113–1122, Lille, France, 07–09 Jul 2015. PMLR.
- P. Whittle. Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society (B)*, pages 143–149, 1980.
- P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.
- H. Wu and X. Liu. Double Thompson sampling for dueling bandits. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances*

- in Neural Information Processing Systems 29*, NIPS, pages 649–657. Curran Associates, Inc., 2016.
- Y. Wu, A. György, and Cs. Szepesvári. Online learning with gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems 28*, NIPS, pages 1360–1368. Curran Associates Inc., 2015.
- Y. Wu, R. Shariff, T. Lattimore, and Cs. Szepesvári. Conservative bandits. In M. Balcan and K. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1254–1262, New York, New York, USA, 20–22 Jun 2016. PMLR.
- Y. Xia, H. Li, T. Qin, N. Yu, and T.-Y. Liu. Thompson sampling for budgeted multi-armed bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI, pages 3960–3966. AAAI Press, 2015. ISBN 978-1-57735-738-4.
- Y. Yao. Some results on the Gittins index for a normal reward process. In *Time Series and Related Topics*, pages 284–294. Institute of Mathematical Statistics, 2006.
- Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1201–1208. ACM, 2009.
- Y. Yue and T. Joachims. Beat the mean bandit. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning*, ICML, pages 241–248, New York, NY, USA, June 2011. ACM.
- Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. In *Conference on Learning Theory*, 2009.
- M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, ICML, pages 928–935. AAAI Press, 2003.
- M. Zoghi, S. Whiteson, R. Munos, and M. Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 10–18, Beijing, China, 22–24 Jun 2014. PMLR.
- M. Zoghi, Z. Karnin, S. Whiteson, and M. Rijke. Copeland dueling bandits. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, NIPS, pages 307–315. Curran Associates, Inc., 2015.
- M. Zoghi, T. Tunys, M. Ghavamzadeh, B. Kveton, Cs. Szepesvári, and Z. Wen. Online learning to rank in stochastic click models. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *PMLR*, pages 4199–4208, 2017.
- S. Zong, H. Ni, K. Sung, R. N. Ke, Z. Wen, and B. Kveton. Cascading bandits for large-scale recommendation problems. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, UAI, 2016.

Index

- D*-optimal design, 242
G-optimal design, 241
 χ^2 -squared distance, 182
 σ -algebra, 21
, 7, 18
a posteriori, 27
a priori, 27
absolutely continuous, 37, 58, 177
action gap, 56
action space, 438
adapted, 26
adversarial bandit, 12, 139, 141
affine hull, 415
anytime, 88
asymptotic optimality, 121, 130, 195, 356, 368, 405, 435
Bachelier-Levy formula, 123
bandits with expert advice, 211
Bayes law, 27
Bayesian regret, 55, 372
Bayesian upper confidence bound, 405
belief space, 458
Bellman optimality equation, 443
Bernoulli bandit, 53, 114
Bernoulli distribution, 32
Bernstein's inequality, 74, 79, 461
best arm identification
 fixed confidence, 361
beta distribution, 378
bias-variance tradeoff, 208
Borel, 22
Borel isomorphism, 44
Borel space, 44
Bregman divergence, 297
Brownian motion, 122, 344
canonical bandit model, 57, 185
cascade model, 346
categorical distribution, 452
Catoni's estimator, 108
cell decomposition, 415
Cesàro sum, 443, 458
Chernoff bound, 126, 355
click model, 346
communicating MDP, 441
concave, 281
conditional entropy, 402
conditional expectation, 32
conditional probability, 26
confidence level, 97
conjugate prior, 377
consistent policy, 193, 271
context, 14
controlled Markov environment, 462
convex optimization, 296
counting measure, 38, 183, 185
covering, 232, 239
Cramer–Chernoff method, 67, 160, 234, 235, 238
cumulant generating function, 134
cumulative distribution function, 32
cumulative generating function, 69
data-dependent bound, 215, 217
degenerate action, 415
descriptive theory, 60
design matrix, 241
deviation matrix, 464
diameter, 441

-
- differential value function, 443
 discount factor, 342
 discounting, 342, 386, 391
 domain of convex function, 280
 dominated action, 415
 doubling trick, 88, 152, 261, 307
 dual norm, 302
 easy partial monitoring problem, 414
 empirical Bernstein, 104, 107
 empirical process, 81
 empirical risk minimization, 217
 entropy, 175, 183, 402
 environment, 10
 epoch-greedy algorithm, 220
 essentially smooth, 283
 essentially strictly convex, 283
 Exp3, 145, 209, 290, 292, 296, 309, 325,
 337, 339, 413
 Exp3-IX, 158, 200, 203, 325, 344
 Exp3.P, 344
 Exp3.S, 344
 Exp4, 213, 338
 expectation, 29
 expectation operator, 39
 exponential family, 114, 133, 135, 197,
 363, 378, 405
 exponential weighting, 145
 exponential weights, 329
 Exponential weights algorithm, 150
 extended real line, 280
 feedback matrix, 412
 Fenchel dual, 281, 329
 filtered probability space, 26
 filtration, 26, 80, 134, 159, 234, 240, 255,
 381
 finite additivity, 21
 first order bound, 162
 first-order optimality condition, 284
 Fisher information, 188
 Fixed Share, 344
 follow the leader, 297, 308
 follow the perturbed leader, 155, 216,
 329, 406
 follow the regularized leader, 296
 Fubini's theorem, 38, 62
 full information, 150, 289, 344
 functional, 39
 fundamental matrix, 464
 gain, 442
 generalized linear model, 228
 Gittins index, 343, 386
 globally observable, 417
 gradient descent, 298
 Gram matrix, 271
 Hahn decomposition, 30
 hard partial monitoring problem, 414
 Hardy–Littlewood, 388
 heavy tailed, 74
 Hedge, 150
 Hellinger distance, 182
 high probability bounds, 157
 history, 10
 Hoeffding's inequality, 80, 355
 Hoeffding's lemma, 74, 127, 150, 240
 Hoeffding–Azuma, 239, 427
 hopeless partial monitoring problem,
 414
 image, 431
 immediate regret, 56
 implicitly normalized forecaster, 312
 importance-weighted estimator, 143,
 144, 162, 217, 304
 independent events, 28
 index, 386
 index policy, 386
 indicator function, 22
 information directed sampling, 407
 instance-dependent bound, 217
 integrable, 30
 Ionescu Tulcea, 59
 Jensen's inequality, 37, 281
 John's theorem, 244
 Jordan curve theorem, 432
 Jordan-Brouwer separation theorem,
 426, 432
 Kahan's algorithm, 145

-
- kernel, 431
 kernel trick, 227
 Kiefer–Wolfowitz Theorem, 242
 Kullback-Leibler divergence, 175
 Laplace’s method, 235
 law of the iterated logarithm, 105, 117, 255
 lazy mirror descent, 307
 learner, 10
 learning rate, 145, 159, 215, 290, 291, 297, 298, 339, 423
 adaptive, 307, 310, 311
 least-squares, 231
 Lebesgue σ -algebra, 36
 Lebesgue integral, 29
 Legendre function, 283, 297, 329
 light tailed, 74
 linear subspace, 431
 link function, 228
 locally observable, 417
 log partition function, 135, 378
 log-concave, 292
 loss matrix, 412
 Markov chain, 423, 440
 Markov kernel, 45
 Markov policy, 440
 Markov process, 49
 Markov property, 457
 martingale, 47, 75, 79, 234
 martingale difference process, 234
 martingale noise, 234
 maximal end-component, 466
 maximal inequality, 117
 maximum end-component, 467
 measurable set, 21
 measurable space, 21
 measure, 21
 median-of-means, 108
 memoryless deterministic policy, 439
 memoryless policy, 439
 metric entropy, 239
 minimax optimal, 170
 minipage, 116
 mirror descent, 150, 296, 329, 339
 MOSS, 114, 116
 multiclass classification with bandit feedback, 218
 multitask bandit, 266, 326
 mutual information, 402
 Nash equilibrium, 308
 negentropy, 283, 299, 327, 339
 neighboring actions, 415
 nonoblivious, 151, 307
 nonparametric, 53
 nonstationary, 151
 nonstationary bandit, 61
 null set, 36
 oblivious, 151, 302, 307
 one-armed bandit, 64, 114, 381
 online gradient descent, 298
 online learning, 16, 217, 257, 296
 online linear optimization, 296
 online-to-confidence set conversion, 256
 operator, 39
 optimal design, 241
 optimal stopping, 382
 optimal value function, 443
 optimization oracle, 216, 333
 optional stopping theorem, 47
 orthogonal complement, 431
 outcome space, 19, 375
 packing, 239
 parameter noise, 316
 parameter space, 375
 parametric, 53
 Pareto optimal, 173
 Pareto optimal action, 415
 partial monitoring, 411
 partially observable Markov Decision Processes, 458
 peeling device, 117
 permutation, 59, 345
 Pinsker’s inequality, 126, 134, 179, 181, 183, 195, 302
 point-locally observable, 433
 polar, 280, 315, 318

-
- policy, 10, 57
 policy iteration, 458
 policy regret, 151
 POMDPs, 458
 position-based model, 346
 potential function, 297
 predictable, 26
 prediction with expert advice, 150
 preimage, 20
 premetric, 126
 prescriptive theory, 60
 prior, 372
 prior variance, 377
 probability kernel, 45, 375
 probability measure, 21
 probability space, 21
 process, 26
 product measure, 38, 58
 product space, 38
 projective, 45
 pseudo regret, 61
 pseudo regret, random, 201
 pushforward, 22
 quadratic variation, 163
 Radon-Nikodym derivative, 37, 58, 62, 185
 random regret, 61
 random variable, 20
 ranked bandit model, 355
 reactive, 151
 reduction, 314, 361
 regret, 10
 - nonstationary, 338
 - tracking, 338
 regret decomposition lemma, 56
 regularizer, 297
 relative entropy, 125, 136, 175, 283
 restless bandit, 343
 retirement policy, 64
 ridge regression, 231
 right stochastic, 423
 right stochastic matrix, 423, 438
 semibandit, 324
 separation oracle, 292, 446
 Sequential Halving algorithm, 366
 signal variance, 377
 signed measure, 21
 similarity function, 211
 simple function, 30
 simple regret, 360
 Snell envelope, 383
 span, 443
 standard optimal stopping, 382
 state space, 438
 static experts, 216
 stationary transition matrix, 442, 464
 stochastic process, 45
 stopping time, 47, 92, 191, 255, 266, 361, 365, 369, 381, 441
 strictly convex, 281
 strongly connected component, 466
 strongly connected MDP, 441
 structured bandit, 54
 sub- σ -algebra, 21
 submartingale, 47
 suboptimality gap, 56
 sufficient statistic, 135, 378
 supermartingale, 47, 238, 255
 supervised learning, 211
 support, 39, 54, 114, 142, 217, 245, 292, 328, 461
 support function, 329, 335
 supporting hyperplane, 285
 Taylor's theorem, 235
 Thompson sampling, 61, 114
 total variation distance, 182
 Track-and-Stop algorithm, 364
 transductive learning, 218
 transition matrix, 442
 trivial events, 41
 trivial partial monitoring problem, 414
 uniform exploration, 360
 union bound, 69
 unnormalized negentropy, 309
 unrealizable, 267, 317
 unstructured bandit, 54

unstructured bandits, 193
Varaiya's algorithm, 392
von Neumann-Morgenstern theorem, 60
weak neighbor, 433
worst case regret, 170