Mathematics & Statistics | Lancaster University

MULTI-ARMED BANDITS:

Sequential Decision Agents in Stochastic Environments

Author:

Lorenzo Croissant

Supervisor:

Dr. Azadeh Khaleghi

Dissertationsubmitted in partial fulfilment for the degree of *Master in Science in Mathematics & Statistics*

April 2018

Acknowledgments

The author wishes for his gratitude to the following people to be publicly known.

First and foremost, the author is grateful to Dr. Azadeh Khaleghi, his supervisor, for her help well beyond this dissertation. Her recommended readings, whether related to this topic or another, provided valuable knowledge and understanding on a wealth of topics. Friendly conversations on a range of machine learning topics in her office strengthened the resolve of the author to pursue it as a career. Here too, her advice proved invaluable and the author is greatly indebted to her.

For her thoughtful advice and regular support throughout his studies, the author thanks Dr. Jenny Wadsworth, his academic advisor.

For her proof-reading and shared passion for typography and elegant language, the author thanks his fellow student, Rachel Bessant, who assisted him many times beyond this thesis.

The author would like to also acknowledge and thank, while not knowing them personally, both Erick C. Montalvan and Andrew R. Dalton. The former's enhancements of the latter's LaTeX dissertation template[1] provided the ground work for the typography of this thesis. As both are public templates, improvements on the template of this thesis will be released in future under a public license, for the benefit of future students.

Finally the author would like to acknowledge Dr. Tor Lattimore and Prof. Csaba Szepesvári for their blog *Bandit Algorithms* which provided the main body of research for this thesis and much of its proofs, theorems and many elements of notation. As this thesis sets out with different goals, its contents have been rearranged, modified, and restyled to what the author deems best. However he is indebted for the quality of their approachable work which greatly contributed to his understanding.

[1] To be found at `https://github.com/ErickChacon/lancs-thesis`.

Contents

0.4pt2pt

Introduction

How often should a recommender system suggest completely new objects in order to map out your preferences? Suppose the recommender only knows whether or not you viewed the specific item it suggested. What then is the cost of this information feedback system relative to a perfect information setting? What is the minimum amount of samples needed to identify the distribution of highest mean amongst a set of unknown distributions? How can one measure the effectiveness of a targeted advertising algorithm? How should one allocate finite resources amongst different projects in order to maximise profits? Why and how are these problems related? Consider the recommender system: it must suggest objects you like as often as possible, to ensure your continued use of the service, while still suggesting new things to attempt to discover new preferences. This is a fundamental trade-off in many problems called the *exploration-exploitation dilemma* which underpins the problems of a field of mathematics known as *bandit theory*. All the questions above can in fact be formulated within this theory, in spite of their apparent differences.

Bandit theory is one of the fields of mathematics blessed with an eyebrow-raising name. The reader might find it amusing that this theory has nothing whatsoever to do with banditry. The seemingly esoteric name comes from a nickname for slot machines, which are sometimes called *one-armed bandits*, for their appearance and the efficiency with which they separate gamblers from their money. A *(multi-armed) bandit problem* consists formally of any sequential allocation problem where an *agent* interacts with an *environment*, receiving *rewards* X_t (a *bandit feedback*) based on its actions. They are generally referred to as *K-armed bandits*, where an agent is faced with a row of K proverbial slot-machines, some of which will have positive pay-off, and where its role becomes to effectively identify and profit from the arms with the highest pay-off. How it carries out this task, its *policy*, is the main object of bandit theory. One seeks to find an optimal policy or class of policies for the specific problem considered, and quantify what this optimality means in practice. There are many different types of bandits within bandit theory. The type of a bandit can be decomposed into two qualitatively different attributes: a paradigm and a specific setting. These are different in that paradigms are mutually exclusive, whereas one can consider the same setting in several paradigms. We will begin by outlining the main paradigms.

The difference in paradigm hinges on the assumptions made about the slot ma-

chine's "randomness". A first approach would be to take usual assumptions of stochastic behaviour of the rewards. For example, consider the sequential clinical trial design problem. Given two new drugs, the traditional way of testing them is to assign to each a fixed number of test subjects and then pick the most effective drug after the trial ends. However applying this method will lead to some unfortunate subjects dying from the lesser drug. In a sequential experiment, in contrast, patients are added in turn to either drug group and their outcome observed before assigning the next patient. The problem of best exploiting this new information framework is considered the first problem in bandit theory. It was proposed by Thompson[34], who provided a Bayesian sampling strategy still studied today for its effectiveness[1]. Thompson assumed that patients' outcomes were independent and that the environment was stationary, a particular set of assumptions called the IID *stochastic bandit*[28], but the fundamental feature of a stochastic bandit is simply that the environment generates stochastic rewards.

There is a special case of the stochastic bandit which stands out and is important to mention: the *Markovian bandit*. Above, it was assumed that the environment was stationary, i.e. that the reward distributions for each arm were the same throughout. In the Markovian bandit, one allows the environment to operate as a stochastic process, with states that may influence the reward distributions. The original problem for which this was designed was in fact the allocation of funds to research projects[15] at regular time intervals. With this in mind, see that the multiple projects (arms) depend on some unknown presumably stochastic process which determines how they progress. This environment is formally called a *Partially Observable Markov Decision Process (POMDP)* which is a general formulation[33] for *reinforcement learning*. While bandits predate the first explicit works in machine learning[29,30], they are a textbook example of reinforcement learning as they learn to make decisions via the rewards they obtain from an unknown environment.

Consider the problem of ad-serving, or how an algorithm should allocate to each viewer of a web-page one ad from its collection in order to maximise the likelihood of the viewer clicking on it. This is a classic problem in bandit theory[10,37], but is not adequately addressed by either stochastic or Markovian bandits. In real world situations, assumptions such as independence are often violated, and interacting with human behaviour is particularly sensitive to this problem. In response, Auer et al.[5] introduced the *adversarial bandit* to remove nearly all the assumptions of stochastic bandits. Here, at each time step, an adversary chooses the rewards for each arm of the bandit completely removing any assumptions about independence or stationarity in rewards. While this could appear to make bandit problems virtually unsolvable, that is not the case. In fact, these problems are not much more difficult than the worst-case stochastic bandit problem one could imagine[7]. Taking the simple assumption that the adversary in this context has a policy of his own, which doesn't

use knowledge of our policy, one can achieve good theoretical results. More importantly, this framework has allowed much better real world results and contributed to the success of bandit theory.

Having outlined all three paradigms, the reader might now expect us to detail all the main researched settings. This is unfortunately not feasible as, in a testament to the flexibility of bandit theory, there are simply too many settings. We will give a very short list of some settings, without detailing their exact framework, as examples, and provide further reading. Each of these settings tends to answer one or more simple real world problems addressed by bandit theory, and improve the achieved performance on these problems. In many problems, for instance, there is some available data that can help the agent improve its decision. This is addressed[21] by *contextual bandits*, which can be used in any paradigm, and can incorporate further restrictions such as linearity or sparsity of the context. Similarly, in many problems the feedback is slightly different from canonical bandit feedback. It may be delayed in time, giving *delayed-feedback bandits*[10], may include some information about other arms than the one played in *bandits with side-observations*[24], or be composed of only comparisons between arms in *duelling bandits*[38].

Another testament to the success of bandit theory is the range of applications covered by contemporary bandit theory. Beyond ad-serving, clinical trials, and project management, multi-armed bandits have shown effectiveness for instance in recommender systems, sometimes using similarity graphs between items to improve performance[35], packet routing[6], search engines[38], and influence maximisation[36]. This is far from covering the range of settings and applications in the literature but should convince the reader that bandit theory reaches far further than the contents of this thesis.

The premise of this thesis shall thus be to outline an approachable informative introduction to bandit theory that would allow a reader with a graduate level in mathematics to build solid basics in bandit theory. Unfortunately, in such a small space the author does not feel that all paradigms nor all settings can be addressed with the level of rigour he considers necessary to sustain the premise of this thesis. The question is then: where to begin? The author believes the answer is to begin with the IID stochastic bandit as introduced by[28] and presented above. There are several reasons for this: it is a simple framework, it is also the first bandit framework to have been presented, but most importantly it is a foundation which introduces the required components to understand other bandit problems. A good understanding of the simple stochastic bandit enlightens an exploration of adversarial or Markovian bandits as one can understand *how* the changes in assumptions lead to changes in the guarantees. The author won't get ahead of himself and give examples here, but the author will give passing remarks once the reader has been made familiar with

the characteristics of the stochastic bandit problem.

Having restricted ourselves to the simple stochastic bandit, let us now explicitly outline what this framework is. An *instance* of the stochastic K-armed bandit problem consists of an agent, which follows a policy, which takes each turn exactly one action in a set of size K, immediately receives a reward for the arm it played and then moves to the next round. We will generally further assume that arms are independent and rewards from the same arm are also independent, that reward information is not altered, and that all arms are playable each round. We will also assume (unless otherwise specified) that rewards are bounded[2]. There are a few implicit assumptions which relate more specifically to the situations where bandits can be used. For instance, note that for the reward distributions to remain the same the overall goal for which the agent was developed must not change either. These are mostly implementation issues, however, and we won't pay them much heed here.

This thesis aims to contribute an approachable introduction to stochastic bandits which can be understood with a graduate background in probability. Thus to cover required material from other fields, we will begin Chapter 1 with a quick outline of foundational material in measure theory which will allow us to redefine probability, random variables, integration and finally density. Then we will do the same for information theory, presenting a quantity of the difference between two distributions, and some inequalities related to it. With these foundational principles we will then in Chapter 2 cover general notions of bandit theory, our measure of performance and concentration inequalities, followed by general results for performance in the regular stochastic bandit problem. These general performance results will allow us to benchmark two classes of algorithms in Chapter 3, with an analysis of their respective regrets. Should the reader require them at any moment, a notational appendix and the bibliography are to be found at the end of this thesis.

[2]This issue will be nuanced in section 2.1.3, by a class of low variance random variables. However, in general we will state results for bandits with rewards bounded in $[0, 1]$. Note that any bounded reward can be rescaled to $[0, 1]$.

Chapter 1

Foundational Preliminaries

RIGOUROUS definitions are an essential prerequisite to any mathematical exercise, and bandit theory is no exception. While not constitutive of the main body of bandit theory, notions outlined in this chapter are required to proceed in an orderly manner in a survey of multi-armed bandits and of the stochastic bandit specifically. This chapter assumes little familiarity with the topics of measure or information theory, neither does it seek to be an introduction to them. It will present only the required components upon which to base the rest of this thesis. We will begin with an exploration of measure theory, defining or redefining the concepts of probability, including integration, density, and measures. We will then immediately apply these building blocks to a succinct brief in Information theory, whose purpose is to present to the reader the Kullback-Leibler divergence and some of its properties.

1.1 Elements in Measure Theory

MEASURE THEORY allows us to reformulate and generalise many statements about probability distributions. This new field refines the theory of probability and allows us to derive a new understanding of what probabilities, events, outcomes, and distributions are. We will assume that the basic notions of sample space and random variables are known to the reader. To go beyond naive probability, we will need to rebuild our definitions from new classes of sets. We will begin by introducing measures, then we will extend them to a key theorem, which will allow us to formulate a generalised probability density, unifying probability mass and density functions, and allowing us to derive new properties of distributions with information theory. Proofs of results will be omitted, but can be found in Leadbetter et al.[23].

1.1.1 Measures

We consider first an arbitrary topological space $\mathcal{X}$, and are interested in $\mathcal{E} \subset \mathcal{P}(\mathcal{X})$, a class of sets on $\mathcal{X}$. We will take the normal operations on sets, and we will denote set-theoretic difference as "$-$" and define some behaviours $\mathcal{E}$ can have.

Definition 1.1.1 (Rings and Fields)**.** A non empty class $\mathcal{E}$ is a *ring* if for all $E, F \in \mathcal{E}$: $E \cup F, E - F \in \mathcal{E}$. Further, a *field* is a ring closed under the complement in $\mathcal{X}$. A ring closed under countable union is called a σ*-ring*, and naturally it is a σ-field or σ-algebra if it is also closed under complements.

In probability we will be mostly interested in σ-algebras, but measure theory includes work on more general cases, like rings[23, p. 23]. A noteworthy example of a σ-algebra is the Borel σ-algebra $\mathcal{B}$ on the real line. Consider the collection of all open intervals on $\mathbb{R}$, or more generally of all open sets in $\mathbb{R}$. The class of Borel sets for the real line is the σ-ring generated by this collection. This can be shown to also be a σ-algebra, which we denote by $\mathcal{B}$. The class $\mathcal{B}$ contains for instance all one-point and countable sets, as well as all intervals in $\mathbb{R}$. Now that we have clearly defined and explored classes of sets, we will set out properties of set functions and define measures.

Definition 1.1.2 (Set functions and their properties)**.** A map M from $\mathcal{E}$ to some set S is a set function if for every $e \in \mathcal{E}$, $M(e) \in S$. We can say that M is:

- Non-negative, if for all $E \in \mathcal{E} : M(E) \geq 0$.
- Countably additive, if for $\{E_i\}_i$ disjoint sets in $\mathcal{E}$, we have $M(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} M(E_i)$.
- Finite, if for all $E \in \mathcal{E} : |M(E)| < \infty$.
- σ-finite, if for all $E \in \mathcal{E}$ there is a sequence of sets $\{E_i\}_i$ in $\mathcal{E}$ with $E \in \cup_{i=1}^{\infty} E_i$ and $|M(E_i)| < \infty$ for all $i \in \mathbb{N}$.
- Real-valued, if for all $E \in \mathcal{E} : M(E) \in \mathbb{R}$.
- Simple, if M is Real valued, $\mathcal{E} = \mathcal{X}$, and $M(E)$ takes a finite number of values.
- Monotone, if for all $E \subset F \in \mathcal{E} \Leftrightarrow M(E) \leq M(F)$.
- Subtractive, if for all $E \subset F \in \mathcal{E}$ such that $E - F \in \mathcal{E}$ and $|M(E)| < \infty$, we have $M(F - E) = M(F) - M(E)$.

Definition 1.1.3 (Measure)**.** A measure μ on $\mathcal{E}$, with $\emptyset \in \mathcal{E}$, is a non-negative, countably additive set-function from $\mathcal{E}$ to $\mathbb{R}$. If $\mu(\mathcal{X}) = 1$ then μ is a *probability measure*.

One of the most important measures is the Lebesgue measure on $\mathcal{B}$ defined simply as the difference between the bounds of the interval, the intuitive length: $\mu((a,b)) = b - a$. The Lebesgue measure further returns the intuitive length, area, and volume when applied to n-dimensional euclidian space.

Definition 1.1.4 (Measurability)**.** Let $(\mathcal{X}, \mathcal{S})$ be a measurable space, i.e. $\mathcal{S}$ is σ-algebra on $\mathcal{X}$, a set E is measurable simply if it is an element of $\mathcal{S}$.

To begin expanding measurability from sets to functions, we introduce the measure space $(\mathcal{X}, \mathcal{S}, \mu)$, where $\mathcal{X}$ is a topological space, $\mathcal{S} \in \mathcal{P}(\mathcal{X})$ is a σ-algebra, and μ is a measure on $\mathcal{S}$. We will also extend our example of the Borel sets to the extended real line $\mathbb{R}^* = \mathbb{R} \cup \{-\infty, \infty\}$. The extended Borel σ-algebra $\mathcal{B}^*$ is defined as $\{B, B \cup \{\infty\}, B \cup \{-\infty\}, B \cup \{-\infty, \infty\} : B \in \mathcal{B}\}$.

Definition 1.1.5 ($\mathcal{S}$-measurability)**.** Let $f : (\mathcal{X}, \mathcal{S}) \mapsto (\mathcal{Y}, \mathcal{T})$ be a set-function between two measurable spaces, f is $\mathcal{S}$-measurable if:

$$\forall E \in \mathcal{T} : f^{-1}(E) \in \mathcal{S}\,.$$

When we work with a real valued set-function f we take $(\mathcal{Y}, \mathcal{T}) := (\mathbb{R}^*, \mathcal{B}^*)$ in this definition.

Notice that the probability measure is a map from the set of all possible events in $\mathcal{X}$, the σ-algebra $\mathcal{S}$, to the interval $[0,1]$, where $\mu(A)$ is 1 if and only if $A = \mathcal{X}$. So a measure acts like the probability distribution of a random variable, in that to each possible combination of outcomes from the sample space $\mathcal{X}$ it associates a probability, and that the measure of the union of all events, akin to the sum of the probabilities of each outcome is equal to 1 by countable additivity. In this subsection we have built only the most basic components of probability theory. In the next subsection we will construct densities using the Radon-Nikodym theorem and measure theoretic integration.

1.1.2 Integration and Density

Before we can introduce the aforementioned theorem, we have to discuss integrability with respect to a measure. Integration is such a key part of probability theory, that it seems only natural to attempt to extend the Riemann integral beyond the real line and to the new classes of sets we have developed. We begin by the case of integrating simple functions, which we can write as a finite sum of indicator functions, where E_i is the sub-class of $\mathcal{E}$ where f takes value a_i. Then we can define the

(Lebesgue) integral of these functions as a simple finite sum:

$$\int f \mathrm{d}\mu = \sum_{i=1}^{n} a_i \mu(E_i) \,.$$

This is an important first step as it can be shown that all non-negative measurable functions are the limit of an increasing sequence $\{f_n\}_n$ of non-negative simple functions. This allows us to formulate a sensible definition for the integral of such functions:

$$\int f \mathrm{d}\mu := \lim_{n\to\infty} \int f_n \mathrm{d}\mu \,.$$

If the integral as such defined is finite, f is integrable. This definition is consistent with the Riemann integral of positive functions on the real line, but it needs to be extended to functions with values in $(-\infty, 0)$. This is done by separating the function into positive and negative parts f_+ and f_-. See now that both are non-negative and if they are integrable we have that f is integrable and its integral takes the finite value:

$$\int f \mathrm{d}\mu := \int f_+ \mathrm{d}\mu - \int f_- \mathrm{d}\mu \,.$$

To proceed to more interesting properties, we will need to define the term "*almost everywhere*" in regards to a statement s. In a fixed measure space, if s holds on all $\mathcal{S}$ except for a set of measure 0, then s is said to hold almost everywhere. In terms of the measure, the collection of points where the property does not hold is negligible, which gives the otherwise ill-defined "almost everywhere" a sensible meaning. Theorem 1.1.1 collects two simple but useful properties of integrable functions.

Theorem 1.1.1 ([23, p. 69]). *If f is integrable with respect to a measure μ, it is finite almost everywhere with respect to μ. Furthermore, if g is measurable, defined almost everywhere and E is a set with 0 measure, then $\int_E g \mathrm{d}\mu = 0$.*

These results are important implicitly in the Radon-Nikodym theorem, but we must clarify a few terminology details before presenting the main result of this subsection.

Definition 1.1.6.

- Absolute continuity: ν is absolutely continuous with respect to μ, denoted $\nu \ll \mu$, if $\mu(E) = 0 \Rightarrow \nu(E) = 0$
- Essentially unique: If f is essentially unique with respect to a property, then any function g with this property is equal to f almost everywhere.

The main result in this section is outlined in Theorem 1.1.2, the Radon-Nikodym theorem. In an intuitive sense it defines a density function whose integral over a

set is the measure of the set. This can be seen as some analog to integrating a density function to obtain an evaluation of the cumulative density function. This result however is much stronger, owing to the use of two measures in the formula, which will allow us to derive in Theorem 1.1.3 a method for changing the measure in an integral.

Theorem 1.1.2 (Radon-Nikodym-Theorem for measures, [23, p. 100]). *Let $(\mathcal{X}, \mathcal{S}, \mu)$ be a σ-finite measurable space, and let ν be a σ-finite measure on $\mathcal{S}$. If $\nu \ll \mu$, then there is an essentially unique finite-valued non-negative measurable function f on $\mathcal{X}$ such that:*

$$\forall E \in \mathcal{S} \,:\, \nu(E) = \int_E f \mathrm{d}\mu \,.$$

Theorem 1.1.3 ([23, p. 102]). *Let μ, ν be σ-finite measures of $(\mathcal{X}, \mathcal{S})$, with $\nu \ll \mu$. If f is a measurable function defined on $\mathcal{X}$ and is either ν-integrable or non-negative, then:*

$$\int f \mathrm{d}\nu = \int f \, \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu \,.$$

In Theorem 1.1.3, $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ is called the Radon-Nikodym derivative of ν with respect to μ. A particular extension of these new derivatives, like in Calculus, is the definition of a "*chain rule*". Theorem 1.1.4 gives this property, which will allow us to write a density as the product of other densities.

Theorem 1.1.4 ("Chain rule" for measures, [23, p. 103]). *Let μ, ν, and λ be σ-finite measures on $(\mathcal{X}, \mathcal{S})$. Then if $\lambda \ll \nu \ll \mu$, we have almost everywhere with respect to μ:*

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\mu} = \frac{\mathrm{d}\lambda}{\mathrm{d}\nu} \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \,.$$

This new formulation of a probability density or mass function, concludes our overview of measure theory. We will use the notation, definitions and results throughout this thesis, and in particular we will immediately apply them in an overview of relevant information theory concepts.

1.2 Elements of Information Theory

EMERGING in the flurry of new research domains to arise in the wake of the Second World War, *information theory* studies the transmission of messages, and the amount of information they contain. We will build upon the origi-

nal purpose of information theory to illustrate its fundamentals. Using the results from section 1.1, we will diverge from the original works of Shannon, and consider applications beyond the transmission of messages. We will ignore the notion of entropy, and present the problem from a hypothesis testing angle, outlined by Kullback[16]. This is because we do not truly need core information theory, but only the parts which transfer to probability theory, in particular the *Kullback-Leibler divergence*. While it is often defined as relative *entropy* in information theory, this connection will not be explored; should the reader like a deeper view in the subject we recommend Cover and Thomas[9], Feinstein[11], or Shannon[31] himself.

1.2.1 Information and Divergence

The fundamental premise of information theory is in signal processing and deals with the analysis of the transmission of messages over *noisy* channels[2]. To by-pass the random scrambling of characters in the channel, the two parties can agree on a common *code*. The source then encodes its message into an object, such as a vector in some vector space, which is transmitted over the channel to the destination's decoder. In the real world, and in signal processing, there is a cost to a more complicated code in terms of the speed of transfer of the message but in applications to statistics we are interested in the point of view of the decoder not the transmission of a message. Instead of a message, say we receive a random value x and we are tasked with determining if it comes from distribution f_1 of f_2. What amount of *information* does x contain about differentiating f_1, f_2? In this thought experiment, taken and adapted from Kullback[16], let us formulate using Bayes' theorem the posterior probability of a hypothesis $i = 1, 2.$ corresponding to x belonging to f_i.:

$$P(H_i|x) = \frac{P(H_i)f_i(x)}{P(H_1)f_1(x) + P(H_2)f_2(x)}\,.$$

We can rearrange the logarithm of the ratio of posteriors for $i = 1, 2$ into a formula that holds almost everywhere with respect to μ:

$$\ln \frac{f_1(x)}{f_2(x)} = \ln \frac{P(H_1|x)}{P(H_2|x)} - \ln \frac{P(H_1)}{P(H_2)}\,.$$

Since the right-hand side is a measure of the change in log-odds between H_1, H_2 before and after x is observed, the left hand side is too, albeit in a less obvious way. We call the left-hand side the information contained in x for differentiating H_1 from H_2. We can then define the mean information using the generalised probability

densities f_1, f_2 for E such that $\mu_1(E) \neq 0$:

$$\begin{aligned} I(1,2;E) &= \frac{1}{\mu_1(E)} \int_E \ln \frac{f_1(x)}{f_2(x)} \mathrm{d}\lambda_1 \\ &= \frac{1}{\mu_1(E)} \int_E f_1(x) \ln \frac{f_1(x)}{f_2(x)} \mathrm{d}\lambda_2 \,. \end{aligned}$$

The second line follows from the Radon-Nikodym derivative $f_1(x) := \frac{\mathrm{d}\lambda_1}{\mathrm{d}\lambda_2}(x)$. Here we notice that the f_i are not important and instead can be fully described by three metrics, λ_1, λ_2, and μ. Thus in the measurable space $(\Omega, \mathcal{F})$ we apply μ to make it a probability space, and two other measures λ_i such that the Radon-Nikodym derivatives of the λ_i with respect to μ are the densities we are trying to separate. Extending our definition from E to Ω:

$$D_{KL}(\lambda_1 \| \lambda_2) := \int \ln \frac{\mathrm{d}\lambda_1}{\mathrm{d}\lambda_2} \mathrm{d}\lambda_1 = \int \frac{\mathrm{d}\lambda_1}{\mathrm{d}\lambda_2} \ln \frac{\mathrm{d}\lambda_1}{\mathrm{d}\lambda_2} \mathrm{d}\lambda_2 \,.$$

This quantity is called the Kullback-Leibler (KL) divergence. Note that if λ_1 is not absolutely continuous with respect to λ_2, the divergence is considered infinite, but otherwise it is finite. This operator is not symmetric and it is therefore not possible to think of the divergence between two measures, but rather from λ_2 to λ_1. The most sensible definition is doubtless the idea which we developed in the simple case as the significance of how helpful information is at separating one measure from the other. We do not need further details about the properties of the Kullback-Leibler divergence, or about other information theoretic concepts such as entropy. As the reader might have guessed, we are mostly interested in using the probability measures of specific random variables, rather than arbitrary ones. Throughout this thesis an important exercise will be the bounding of random variables, and for one such bound we will require a class of results known as the *Pinsker-type inequalities*.

1.2.2 Pinsker-type Inequalities

One reason we have developed all these precise tools, is to form an acceptable background for establishing and proving specific results. Our preeminent concern will be the discovery of bounding inequalities. An important bounding equality related to Kullback-Leibler divergence is the Pinsker inequality. For two measures in the measurable space $(\Omega, \mathcal{F})$ as before, we define the total variation distance $\delta(\lambda_1, \lambda_2) = \sup\{|\lambda_1(E) - \lambda_2(E)| : E \in \mathcal{F}\}$. Then the Pinsker inequality states that:

$$2\delta(\lambda_1, \lambda_2)^2 \leq D_{KL}(\lambda_1 \| \lambda_2) \,.$$

This subsection consists in the derivation of a bound for the sum of measures of two complementary events, which we will refer to as the *Pinsker-type* inequality as it contains the Kullback-Leibler divergence like the true Pinsker inequality. This identity is given in Theorem 1.2.1 and is proven immediately afterwards.

Theorem 1.2.1 (Pinsker-type inequality[20]). *Let λ_1, λ_2 be probability measures on $(\Omega, \mathcal{F})$. For all $E \in \mathcal{F}$, we have:*

$$\lambda_1(E) + \lambda_2(E^c) \geq \frac{1}{2}\exp(-D_{KL}(\lambda_1 \| \lambda_2))\,.$$

Proof. Consider a probability space $(\Omega, \mathcal{F}, \mu)$, let λ_1 and λ_2 be two further measures on this space, respectively with Radon-Nikodym derivatives f_1 and f_2 with respect to μ. For $E \in \mathcal{F}$, we begin by reducing the left hand side to an integral of a minimum on Ω. See that:

$$\begin{aligned}
\lambda_1(E) + \lambda_2(E^c) &= \int_E f_1 \mathrm{d}\mu + \int_{E^c} f_2 \mathrm{d}\mu \\
&\geq \int_E \min(f_1, f_2)\mathrm{d}\mu + \int_{E^c} \min(f_1, f_2)\mathrm{d}\mu \\
&\geq \int \min(f_1, f_2)\mathrm{d}\mu\,.
\end{aligned}$$

Note now that $f_1 + f_2 = \max(f_1, f_2) + \min(f_1, f_2)$. Using the unit integrability of measures this trivial fact allows us to derive the much more useful statement that $\int \max(f_1, f_2)\mathrm{d}\mu \leq 2 - \int \min(f_1, f_2)\mathrm{d}\mu \leq 2$. Now:

$$\begin{aligned}
\int \min(f_1, f_2) &\geq \frac{1}{2}\int \min(f_1, f_2)\mathrm{d}\mu \int \max(f_1, f_2)\mathrm{d}\mu \\
&\geq \frac{1}{2}\int f_1 \mathrm{d}\mu \int f_2 \mathrm{d}\mu \\
&\geq \frac{1}{2}\int \left(\sqrt{f_1 f_2}\right)^2 \mathrm{d}\mu \\
&\geq \frac{1}{2}\left(\int \sqrt{f_1 f_2}\mathrm{d}\mu\right)^2 \\
&\geq \frac{1}{2}\exp\left(2\ln \int \sqrt{f_1 f_2}\mathrm{d}\mu\right) \\
&\geq \frac{1}{2}\exp\left(2\ln \int f_1 \sqrt{\frac{f_2}{f_1}}\mathrm{d}\mu\right)\,.
\end{aligned}$$

Further as f_2 is absolutely continuous with respect to f_2, we know that $f_1 > 0$ implies $f_1 f_2 > 0$. This will allow us to apply Jensen's inequality:

$$\begin{aligned}\lambda_1(E) + \lambda_2(E^c) &\geq \frac{1}{2}\exp\left(2\int f_1 \ln\sqrt{\frac{f_2}{f_1}}\mathrm{d}\mu\right) \\ &\geq \frac{1}{2}\exp\left(-\int f_1 \ln\frac{f_1}{f_2}\mathrm{d}\mu\right) .\end{aligned}$$

Replacing f_1 and f_2 by their definition as Radon-Nikodym derivatives yields the result:

$$\lambda_1(E) + \lambda_2(E^c) \geq \frac{1}{2}\exp\left(-D_{KL}(\lambda_1\|\lambda_2)\right) .$$

□

Throughout this chapter, we have reformulated probability theory in terms of topological spaces, σ-algebras and measures. By redefining functions on measurable spaces we developed a formalism which is consistent with the axioms of probability, but provides us with new flexibility. This new framework led us to new insights about the definition of densities, unifying the countable and uncountable domains. Further, we derived a new calculus for measures, using the Radon-Nikodym theorem, and Lebesgue integration, giving rise in the measure theoretic framework to new concepts of continuity and uniqueness. These ideas will be used throughout this thesis, and notably we applied them straight away to a probabilistic exploration of information theoretic concepts related to the Kullback-Leibler divergence, which is used throughout the field of machine learning. While quantifying the information distinguishing one distribution from another, it also allows us to derive a family of bounds, including the Pinsker inequality and Theorem 1.2.1.

Chapter 2

Stochastic Bandits

MODELLING complex, partially unknown real world systems is typically done by replacing unknown mechanics with random approximations. This is the case of the real world one-armed bandit for instance: one gambles against a seemingly random slot machine which in fact is wholly deterministic. It is therefore natural to study first the problem of sequential action allocation in a stochastic environment. In this chapter, we will establish some mathematical foundations which we will use to evaluate the performance of algorithms in chapter 3. We will begin by outlining mathematical concepts which will be needed to study the performance of bandit algorithms, and then by characterising the general properties of the stochastic bandit problem.

2.1 Mathematical notes

THE main quantity of mathematical interest in this thesis is the concept of regret. Like in all other machine learning problems, bandits are motivated by the optimisation of a loss function, equivalent to the maximisation of obtained rewards. As the rewards are random variables, it will be natural to consider the expectation of accumulated rewards as a function of time. The question is what to compare rewards to, in order to obtain a meaningful loss which the algorithm can learn from. This will motivate our definitions of regret, which we will use to show a quintessential theorem called the regret decomposition identity. Later, we will turn our attention to the problem of concentration inequalities on the mean of certain random variables. We will define sub-gaussianity, then show two important concentration inequalities, which will allow us to bound the probability of over or underestimating by the sample mean of the true mean. Before anything else, we will define the bandit problem and outline notation.

2.1.1 The Bandit Problem

In the introduction we outlined briefly what a bandit problem is on a high-level, now we will take a more mathematically rigorous definition of all terms we will use throughout this thesis. As the reader will recall, an *instance* ν of the bandit problem (the environment) consists of a set $\mathcal{K}$ of K arms, with distributions $\{P_i\}_{i\in\mathcal{K}}$. These distributions will be assumed to have hidden parameters θ_i, which determines their mean μ_i. We aren't trying to do inference on the P_i, thus we won't use θ_i in practice, but there are a range of results in the literature[8,17] where this does become important, hence we have brought it to the reader's attention. The bandit will last a natural number N, the *horizon*, of turns which we will index with the interval $[\![N]\!] := \{1, 2, \dots, N\}$. We will let a time index t or n run over this interval, so that at each time t the agent will choose an action $I_t \in \mathcal{K}$ according to a policy and receive a real valued reward X_t from the environment. In our IID stochastic bandit, X_t will be randomly drawn from P_{I_t}, independently of all other rewards.

The particularity of bandit problems is that information from the *history* of actions and rewards allows the agent to make informed decisions. Before we can properly define a policy then, we must first formalise histories. The history at time t, H_t is the set of all past actions and rewards. Since these are in general random, we define:

$$H_t := (I_1, X_1, \dots, I_{t-1}, X_{t-1}) \text{ for } 1 < t \leq N \text{ and } H_1 := \emptyset\,.$$

Since $I_t \in \mathcal{K}$ and $X_t \in \mathbb{R}$, the reader will easily see that since the space of all possible realisations of H_t is the cartesian product of t copies of $\mathcal{K} \times \mathbb{R}$. We will write this product using powers to signify the repeated cartesian product so that $H_t \in (\mathcal{K} \times \mathbb{R})^t$. The history is a random variable, which allows us to create a filtration of the $\mathcal{H}_t$, the σ-algebras generated by H_t, for all $t \leq N$:

$$\mathscr{H} := \{\mathcal{H}_t\}_{t=1}^N = \{\sigma(H_t)\}_{t=1}^N\,.$$

Now that histories have been defined we can move on to policies, which are formally any $\mathscr{H}$-adapted mapping from $(\mathcal{K} \times \mathbb{R})^N$ to $\mathcal{K}$. Less formally, this means that a policy at each time t maps the history it has available (h_t a realisation of H_t) to one arm in $\mathcal{K}$. This is what we colloquially refer to as the agent "choosing" an action. The $\mathscr{H}$-adaptability property simply requires that the policy choose based only on information it has already collected and can not "see ahead". We will also require that π be oblivious to the parameters θ_i. We will denote by $\pi(t)$ the action taken at time t by π, implicitly given H_t. We group together all such instances by their number of arms K into the classes $\mathcal{E}_K$, and let $\mathcal{E}$ be the class of all these instances regardless of their number of arms.

2.1.2 Regret Properties

What is referred to as the *regret* can be confusing, as such we will begin by clarifying its exact definitions used in this thesis. The regret is a random variable, R_n, defined[7] as the difference between repeating the action of highest reward and the received sum of rewards from actions I_t. In other words, the regret represents the loss in the bandit setting relative to a perfect information setting. Mathematically:

$$R_n = \max_{i \in \mathcal{K}} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t} \, .$$

It would be natural to examine the *expected regret* $\mathbb{E}[R_n]$, where the agent competes against the expected value of the maximal reward obtainable by playing the same arm repeatedly. In practice however, we tend to examine the weaker[1] notion of *pseudo-regret*, as defined in Definition 2.1.1, where one competes only against the sequence which is optimal in expectation. The pseudo-regret is colloquially referred to as the regret in literature, for simplicity, and will sometimes henceforth be referred by this simplification. The reader should, however, bear in mind this semantic simplification.

Definition 2.1.1 (Pseudo-Regret)**.** In a stochastic bandit problem with K arms, for round $0 < n \leq N$, we define the pseudo-regret as:

$$\bar{R}_n = \max_{i \in \mathcal{K}} \left\{ \mathbb{E} \left[\sum_{t=1}^{n} X_{i,t} - X_{I_t,t} \right] \right\} \, .$$

It is straightforward to reorder this definition into a more tractable form by defining $\mu^* = \max_{i \in \mathcal{K}} \{\mu_i\}$ to be the expected payoff of the optimal arm.

Corollary 2.1.1 (Tractable pseudo-regret)**.**

$$\bar{R}_n = n\mu^* - \mathbb{E} \left[\sum_{i=1}^{n} X_{I_t,t} \right] \, .$$

From these simple principles, we can now introduce the main result of this section, whose proof will be immediately given thereafter.

Theorem 2.1.2 (Regret Decomposition Identity)**.** *In a stochastic bandit with K arms, for* $0 < n \leq N$*, let* $\Delta_k = \mu^* - \mu_k$ *be the sub-optimality gap for arm k, and let* $T_k(n) :=$

[1] Indeed, Definition 2.1.1 will show that the expected regret is always greater.

$\sum_{t=1}^{n} \mathbb{I}\{I_t = k\}$ *be the random variable counting the number of times arm* k *is chosen in* n *rounds. We have:*

$$\bar{R}_n = \sum_{k \in \mathcal{K}} \Delta_k \mathbb{E}[T_k(n)] \,.$$

Proof. This simple proof relies on re-indexing the sum in the regret from the number of rounds to the number of arms:

$$\begin{aligned}
\bar{R}_n &= \sum_{t=1}^{n} \mu^* - \sum_{t=1}^{n} \mu_i \\
&= \sum_{t=1}^{n} (\mu^* - \mathbb{E}[X_{I_t,t}]) \\
&= \sum_{t=1}^{n} \mathbb{E}[\Delta_{I_t}] \\
&= \sum_{t=1}^{n} \sum_{k \in \mathcal{K}} \mathbb{E}[\Delta_k \mathbb{I}\{I_t = k\}] \\
&= \sum_{k \in \mathcal{K}} \Delta_k \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\{I_t = k\}\right] \\
&= \sum_{k \in \mathcal{K}} \Delta_k \mathbb{E}[T_k(n)] \,.
\end{aligned}$$

□

Theorem 2.1.2 concludes the list of regret properties we need to review, and we will now move on to a discussion of the estimation of a mean, which will allow us to derive tail probability bounds we will need.

2.1.3 Estimation of a Mean

In order to set up algorithms for stochastic bandits it is necessary to effectively estimate the mean pay-off of each arm. Indeed, each time we pull a sub-optimal arm to refine its mean, we incur a penalty with expectation Δ_k. This seems like a straightforward task, after-all there is an unbiased estimator for the mean, the sample mean, which we can adapt. However, being unbiased is not the paramount property in the case of a bandit problem. If the estimator is unbiased but has high variance it will be difficult to determine which arm has the highest true mean. Recall that the variance (i.e. inacuracy) of the sample mean is $\frac{\sigma^2}{n}$, where n is the number of samples and σ^2 is the variance of their underlying distribution.

We would like to define a framework to describe the distribution of $\hat{\mu}$, which is unknown. To do so we will look at the tail probabilities $P(\hat{\mu} \geq \mu + \epsilon)$ and $P(\hat{\mu} \leq \mu - \epsilon)$ and attempt to bound them. We could use Chebyshev's inequality or the central limit theorem, but the first one is a weak bound and the second one is asymptotic. Instead we will define a new property of a random variable which will allow us to derive new properties about its tail probabilities.

Definition 2.1.2 (Sub-gaussianity)**.** A random variable X is σ^2-sub-gaussian[2] if it satisfies, for all $\lambda \in \mathbb{R}$:

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

Intuitively X is sub-gaussian if its tails are lighter than the gaussian distribution, which is to say that it has lower tail probabilities. We follow this with some simple results about independent sub-gaussian random variables.

Lemma 2.1.3 (Properties of sub-gaussian random variables[18])**.** *Let X be a sub-gaussian random variable.*

i. We have $\mathbb{E}(X) = 0$ and $\operatorname{var}(X) \leq \sigma^2$.

ii. For $c \in \mathbb{R}$, cX is $c^2\sigma^2$-sub-gaussian.

iii. For $X_1 \perp X_2$ σ_1-sub-gaussian, and σ_2-sub-gaussian respectively, $X_1 + X_2$ is $(\sigma_1^2 + \sigma_2^2)$-sub-gaussian.

Proof.

i. We consider $\lambda \neq 0$, as this is a vacuous case. Note that we can expand the definition as

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$
$$\mathbb{E}\left[\sum_{n=0}^{\infty}\frac{(\lambda X)^n}{n!}\right] \leq \sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{\lambda^2\sigma^2}{2}\right)^n.$$

By a Taylor expansion[27] around 0, we obtain for all $\lambda \in \mathbb{R}$, as $t \to 0$:

$$\lambda\mathbb{E}[X] + \lambda^2\mathbb{E}[X^2] \leq \frac{\lambda^2\sigma^2}{2} + R_2(\lambda).$$

[2] In the Gaussian case there is equality, hence the name.

We separate cases where $\lambda > 0$ and $\lambda < 0$, and divide by λ. Taking the limit to 0 gives:

$$E(X) \geq 0 \text{ if } \lambda < 0 \text{ and } E(X) \leq 0 \text{ if } \lambda > 0 \Rightarrow E(X) = 0\,.$$

For the variance we divide instead by λ^2, and take the limit as $\lambda \to 0$ to remove $R_2(\lambda)$:

$$\text{var}\,(X) = \mathbb{E}[X^2] \leq \sigma^2\,.$$

ii. Letting $\lambda' = \lambda c \ \in \mathbb{R}$ in the definition gives the result.

iii. A simple computation gives:

$$\begin{aligned}\mathbb{E}[\exp(\lambda(X_1 + X_2))] &= \mathbb{E}[\exp(\lambda X_1)]\mathbb{E}[\exp(\lambda X_2)] \\ &\leq \exp\left(\frac{\lambda\sigma_1^2}{2}\right)\exp\left(\frac{\lambda\sigma_2^2}{2}\right) \\ &= \exp\left(\frac{\lambda(\sigma_1^2+\sigma_2^2)}{2}\right)\,.\end{aligned}$$

Thus $X_1 + X_2$ is $(\sigma_1^2 + \sigma_2^2)$-sub-gaussian. □

To formalise our intuition of the lightness of tails of sub-gaussian random variables we introduce the following concentration inequality, which we prove using Chernoff's method.

Theorem 2.1.4 (Concentration of Sub-gaussian Random Variables[18])**.** *If X is a σ^2-sub-gaussian, then $P(X \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$.*

Proof. Let $\lambda > 0$. As exponentiation conserves inequalities we have:

$$P(|X| \geq \epsilon) = P(\exp(\lambda|X|) \geq \exp(\lambda\epsilon))\,.$$

As $\lambda\epsilon > 0$, we can apply Markov's inequality to $|X|$, which gives us an upper bound for the above:

$$\begin{aligned}P(X \geq \epsilon) \leq P(|X| \geq \epsilon) &\leq \mathbb{E}[\exp(\lambda X)]\exp(-\lambda\epsilon) \\ &\leq \exp\left(\frac{\lambda^2\sigma^2}{2} - \lambda\epsilon\right)\,.\end{aligned}$$

Now we choose λ to minimise this bound as it holds for all $\lambda > 0$. Observe that we

take $\lambda = \frac{\epsilon}{\sigma^2}$ and thus have:

$$P(X \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

□

From theoreom 2.1.4, we can derive a bound for the sample mean we were interested in.

Corollary 2.1.5 (Hoeffding's bound)**.** *Let $X_i - \mu$ be independent σ^2-sub-gaussian random variables. Then $\hat{\mu}$ has tail probability bounds:*

$$P(\hat{\mu} \geq \mu + \epsilon) \leq \exp\left(-\frac{n\epsilon}{2\sigma^2}\right) \text{ and } P(\hat{\mu} \leq \mu - \epsilon) \leq \exp\left(-\frac{n\epsilon}{2\sigma^2}\right).$$

This concludes our discussion of estimation of means as we have found satisfactory bounds on tail probabilities for the sample mean. We now outline the general characteristics of the stochastic bandit problem in terms of the regret properties which we will use to compare competing algorithms.

2.2 Characterising the Stochastic Bandit Problem

AFTER these basic concepts, we will now move to the heart of the stochastic bandit problem. In the previous section, we defined the bandit problem, the regret, and related ideas; now, in this section we will show what behaviour the regret can have in a bandit problem. The results here are not about *a* policy or *an* algorithm, but fundamental characteristics of the stochastic bandit problem and in particular of what the optimal regret behaviour is. First we will analyse optimal asymptotic behaviour, the best that can be achieved over any instance, followed by the best achievable finite-time regret on the most difficult instance. These two complimentary bounds will allow us to evaluate every facet of the performance of algorithms when we present them in the following chapter.

2.2.1 Optimal Asymptotic Regret Growth

The main result in this section is Theorem 2.2.1, which was presented in Lai and Robbins[17] and is one of the foundational results in bandit theory. The result is quite simple, it states that the regret is asymptotically $\Omega(\ln(n))$, and gives a lower bound on the constant term of this logarithmic growth. The main objective of this section

is to convey to the reader understanding of the complicated original proof. To do so we will sketch the arguments of Lai and Robbins, which hold for any distribution in the exponential family, as we go about proving the simpler case of a bandit with Bernoulli distributed rewards in a proof adapted from Bubeck et al.[7]. The First step is to introduce the class of *consistent* policies for which the theorem holds.

Definition 2.2.1 (Consistency)**.** A policy π is consistent if for any stochastic bandit $\nu \in \mathcal{E}$, and for all $\alpha > 0$, we have that:

$$R_n^\nu(\pi) = \mathcal{O}(n^\alpha) \text{ as } n \to \infty\,.$$

Bubeck et al. define them as policies where $\mathbb{E}[T_i(n)] = o(n^\alpha)$, which the reader will note is the same. What is important is that these policies pick sub-optimal arms with a sub-polynomial growth. It may seem very restrictive to the reader to choose such a condition, but it is in fact not. This is slightly technical and we should not get ahead of ourselves, but the jist of it is that only sub-linear policies are applicable in practice, so we don't loose anything by ignoring policies of higher regret order. The flaw in this argument is that it relies on the supposition sub-linear policies exist, which we haven't shown yet. They do exist, we will see some in the next chapter where we will also give example of sub-polynomial policies, which justifies our restriction to consistent policies here.

We are now ready to introduce Theorem 2.2.1, and dive intro the proof, but there are a few caveats the author wishes to raise first. This theorem relies on some hidden assumptions which apply only to single parameter distributions from the exponential family which are bounded. There are extensions[8] to this result that make it valid in more situations and explain why these details are generally not presented for simplicity. Second caveat: this theorem is an asymptotic *lower bound*, there is no guarantee for an arbitrary policy that it will be asymptotically logarithmic. Likewise, there is no guarantee that in finite time we will reach the logarithmic domain[14] even for an optimal policy. Third, this is a lower-bound but it is tight, as we will see when we give a policy that attains this bound in the next chapter.

Theorem 2.2.1 (Assymptotic Growth Lower-Bound, [17, thm. 1])**.** *For all consistent policies π, for all instances $\nu \in \mathcal{E}$, we have:*

$$\liminf_{n\to\infty} \frac{R_n^\nu(\pi)}{\ln(n)} \geq \sum_{i:\Delta_i>0} \frac{\Delta_i}{D_{KL}(P_i \| P_{i^*})} \text{ for } p_1 < p_2\,.$$

Proof. This proof technically proves Corollary 2.2.3, but will outline the proof of Theorem 2.2.1 so that the reader can understand how this simple and often given proof extends to the more general case. In this proof, we restrict ourselves to a bandit with

Bernoulli reward distributions. We will first set up notation and conditions, creating two similar instances, then we will outline the rest of the proof.

One of the hidden conditions of Theorem 2.2.1 is that the Kullback-Leibler divergence must be continuous over the parameter space of the arms. Here the Bernoulli bandit simplifies things a lot, as for two bernoulli with parameters p_1,p_2 we have:

$$D_{KL}(p_1\|p_2) = p_1 \ln\left(\frac{p_1}{p_2}\right) + (1-p_1)\ln\left(\frac{1-p_1}{1-p_2}\right),$$

And this is easily seen to be continuous on $(0,1)$. Let ν and ν' be two instance in $\mathcal{E}_K$. Without loss of generality, let arm 1 be optimal in ν with a mean of μ_1. We design ν' to have optimal arm 2 with mean μ_2' and all other arms identical to ν. Since we required D_{KL} to be continuous, we can choose $\mu_1 < \mu_2' < 1$ such that for $\epsilon > 0$:

$$D_{KL}(\mu_2\|\mu_2') \leq (1+\epsilon)D_{KL}(\mu_2\|\mu_1).$$

Since we have given this peculiar set-up without explanation, and we have not yet started the proof itself we shall take a step back and explain the outline of the proof. About these two bandits, first: the continuity condition guarantees that we can make the divergences of these two instance arbitrarily close. Then, we want to show that this makes the policy unable to distinguish the two instances, meaning it has the almost the same behaviour on both. Here the assumption of consistency comes in to bound the number of times our policy mistakenly plays arm 1 in ν', $\mathbb{E}_{\nu'}[T_2(n)]$. Which then also lower bounds how many times it will have played arm 2 in ν where it is sub-optimal, $\mathbb{E}_{\nu}[T_2(n)]$. We will only do this with one arm, but it can simply be repeated for each other sub-optimal arm to obtain the same bound. Then combining this with the regret decomposition identity will give the result.

In practice to link the two instances we will use the empirical estimate of the KL-divergence and a change-of-measure identity. First, we change the way we index rewards so that $X_{2,i}$ denotes the i^{th} pull of arm 2 and not the overall i^{th} pull. Then we define the estimate of the divergence from μ_2' to μ_2 after s pulls of arm 2 as:

$$\hat{d}(s) = \hat{D}_{KL}(\mu_2, \mu_2') = \sum_{t=1}^{s} \ln\left(\frac{\mu_2 X_{2,t} + (1-\mu_2)(1-X_{2,t})}{\mu_2' X_{2,t} + (1-\mu_2')(1-X_{2,t})}\right).$$

This estimate is used in C_n the event which links the two instance:

$$C_n := \left\{ T_2(n) < \frac{(1-\epsilon)\ln(n)}{D_{KL}(\mu_2\|\mu_2')}), \hat{d}(T_2(n)) \leq \left(1 - \frac{\epsilon}{2}\ln(n)\right) \right\}.$$

The first part of the proof will show that C_n has vanishing probability in ν. Define measures $\mathbb{P}_\nu$ and $\mathbb{P}_{\nu'}$ over their respective instances. Then any event in $\sigma(\{X_{2,i}\}_i)$ has the following change of measure from ν to ν':

$$\mathbb{P}_{\nu'} = \mathbb{E}_\nu[\mathbb{I}\{A\}\exp(-\hat{d}(T_2(n)))]\,.$$

In particular for C_n we can lower bound this using the second part of the event:

$$\mathbb{P}_{\nu'} = \mathbb{E}[\mathbb{I}\{C_n\}\exp(-\hat{d}(T_2(n)))] \geq \mathbb{P}_\nu(C_n)\exp\left(-(1-\frac{\epsilon}{2})\ln(n)\right)\,. \tag{2.1}$$

Next we will upper bound $\mathbb{P}(C_n)$ until we can use consistency to show it to be vanishing, to avoid clutter we denote:

$$f_n := \frac{(1-\epsilon)\ln(n)}{D_{KL}(\mu_2\|\mu_2')}\,.$$

Inverting Equation 2.1 and after that using the first part of the definition of C_n:

$$\begin{aligned}\mathbb{P}_\nu(C_n) &\leq n^{(1-\frac{\epsilon}{2})}\mathbb{P}_{\nu'}(C_n)\\ &\leq n^{(1-\frac{\epsilon}{2})}\mathbb{P}_{\nu'}(T_2(n) < f_n)\\ &\leq n^{(1-\frac{\epsilon}{2})}\frac{\mathbb{E}_{\nu'}[n-T_2(n)]}{n-f_n}\,.\end{aligned}$$

The last step follows trivially from $\mathbb{P}_{\nu'}(T_2(n) < f_n) = \mathbb{P}_{\nu'}(n - T_2(n) > n - f_n) = \mathbb{E}[\mathbb{I}\{n - T_2(n) > 1\}]/(n - f_n)$. Now we can use consistency to see that the right hand side above is $o(1)$ as $\mathbb{E}[n - T_2(n)] = o(n^{1-\alpha})$. Now we only need to show that the first part of C_n is vanishing, i.e. $\mathbb{P}_\nu(T_2(n) < f_n) = o(1)$. Since $\mathbb{P}_\nu(C_n) = o(1)$ we need only show that the second part is asymptotically certain. To do so, first note:

$$\mathbb{P}_\nu(C_n) \geq \mathbb{P}_\nu\left(T_2(n) < f_n, \max_{s\leq f_n}\hat{d}(s) \leq \left(1-\frac{\epsilon}{2}\right)\ln(n)\right)\,.$$

Now note that $\hat{d}(s) = \sum Y$ for IID Ys with mean $D_{KL}(\mu_2\|\mu_2')$[3] Thus, by the maximal strong law of large numbers, $\max\hat{d}(s) \to D_{KL}(\mu_2\|\mu_2')$ almost surely as $n \to \infty$. Thus, multiplying both sides of the second part of C_n by $D_{KL}(\mu_2\|\mu_2')/((1-\epsilon)\ln(n))$, we now have:

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{D_{KL}(\mu_2\|\mu_2')}{(1-\epsilon)\ln(n)}\max_{s\leq f_n}\hat{d}(s) \leq \frac{\left(1-\frac{\epsilon}{2}\right)D_{KL}(\mu_2\|\mu_2')}{1-\epsilon}\right) = 1\,.$$

[3]This can be checked by the definitions of $\hat{d}(s)$ and D_{KL} and the law of total probability.

Thus, $\mathbb{P}_\nu(T_2(n) < f_n) = o(1)$ □

To complete this section, we present two simple corollaries applying Theorem 2.2.1 to two common cases, first the case of a multi-armed bandit where arms have gaussian reward distribution, and the second to a two-armed bandit, with Bernoulli rewards.

Corollary 2.2.2 ([19]). *Given two gaussian distributions with variance* 1 *and means* μ *and* $\mu + \lambda$*, their Kullback-Leibler divergence is* $\frac{\lambda^2}{2}$*. Choosing* $\lambda = \Delta_i$ *to maximise the bound, we have from Theorem 2.2.1:*

$$\liminf_{n\to\infty} \frac{R_n^\nu(\pi)}{\ln(n)} \geq \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}.$$

Corollary 2.2.3 ([7, thm. 2.2]). *In a stochastic bandit, with rewards from Bernoulli distributions, the bound from Theorem 2.2.1 becomes by Pinsker's inequality:*

$$\liminf_{n\to\infty} \frac{R_n^\nu(\pi)}{\ln(n)} \geq \sum_{i:\Delta_i>} \frac{1}{2\Delta_i}.$$

As our first important result in the analysis of the stochastic bandit problem, we derived the optimal asymptotic growth rate for any policy. This will allow us to compare limiting performance upper bounds for our algorithms to a benchmark. This proof was also our first intense foray into the details for which we developed tools in chapter 1. The general case proof is more complex[17?], but it is rewarding in the elegance and generality of its result; we do encourage the reader to use this proof as a springboard to understand the general one. In the next section we will give a finite time regret benchmark.

2.2.2 Minimax Regret bound

The *minimax* regret is a fundamental quantity of the difficulty of the stochastic bandit problem. A term borrowed from game theory, it represents the smallest regret achievable in the worst case problem by any policy or algorithm. If the particular instance in $\mathcal{E}$ we are working with is reasonably well behaved, we can achieve lower regret, but if we are in the worst possible instance, the minimax is a meaningful lower bound for the best possible performance of any algorithm. After proving a two lemmas we will compute a lower bound on the minimax, which will prove to be tight.

Recall from Subsection 2.1.1, a single realisation of H_n gives us the entirety of the contents of an n round run against a K-armed bandit. We take the following measurable space in the rest of this section: $((\mathcal{K} \times \mathbb{R})^n, \mathcal{H}_n)$, which is referred to as the *canonical bandit*. The next logical step is to look at measures on this space. Consider the random variables A_t, X_t for $t \leq n$, see that for all $h_n \in (\mathcal{K} \times \mathbb{R})^n$, trivially $A_t(h_n) = a_t$ and $X_t(h_n) = x_t$. This means that in the canonical bandit model, these are fixed and no longer random. More interestingly, they do not depend on an instance or policy. We now set out to prove the aforementioned two lemmas, 2.2.4 and then 2.2.5, which will be used in the proof of our main result on the minimax regret. First, let $\mathbb{P}_\nu$ be the *canonical bandit measure* on $((\mathcal{K} \times \mathbb{R})^n, \mathcal{H}_n)$. This overarching measure will be decomposed immediately in Lemma 2.2.4.

Lemma 2.2.4 ([20]). *In a K-armed bandit instance ν, with arms with distribution P_i for $i \in [\![K]\!]$, we have:*

$$d\mathbb{P}_\nu(h_n) = \prod_{t=1}^{n} P(a_t | a_s, x_s : s \leq t, \pi) \mathrm{d}P_{a_t}(x_t) \mathrm{d}\rho(a_t) \,.$$

Where $P(a_t | a_s, x_s : s \leq t)$ denotes the probability that under policy π the next action taken given the history up to $t-1$ is a_t, and $\rho(a_t)$ is the counting measure[4].

Proof. Expanding, by conditioning, we have:

$$\begin{aligned}
\mathrm{d}\mathbb{P}_\nu(h_n) &= \mathrm{d}\mathbb{P}_\nu(h_n | h_{n-1}) \mathrm{d}\mathbb{P}_\nu(h_{n-1}) \\
&= \prod_{t=1}^{n} \mathrm{d}\mathbb{P}_\nu(h_t | h_s : s < t) \\
&= \prod_{t=1}^{n} \mathrm{d}A_t(h_t | h_s : s < t) \mathrm{d}X_t(h_t | h_s : s < t) \\
&= \prod_{t=1}^{n} \mathrm{d}P(a_t | a_s, x_s : s < t, \pi) \mathrm{d}P_{a_t}(x_t) \mathrm{d}\rho(a_t) \,.
\end{aligned}$$

The last step follows by defining the $\mathrm{d}P(\cdot)$ as the Radon-Nikodym derivatives with respect to an arbitrary measure, in this case, $\rho(a_t)$, the counting measure of $[\![K]\!]$. The proof is complete but we can go one step further and expand the Radon-Nikodym derivative $p_{a_t}$of the $\mathrm{d}P_{a_t}$ with respect to an arbitrary measure λ which dominates all of them, to obtain:

$$\mathrm{d}\mathbb{P}_\nu(h_n) = \prod_{t=1}^{n} \mathrm{d}P(a_t | a_s, x_s : s < t, \pi) \mathrm{d}\rho(a_t) \mathrm{d}\lambda(x_t) p_{a_t}(x_t) \,.$$

[4] We could use any arbitrary measure here, but we choose to take the counting measure as $a_t \in [\![K]\!]$.

□

Lemma 2.2.5 ([20]). *For two instances $\nu, \nu' \in \mathcal{E}_K$, with arms P_i and P'_i respectively, under the same policy π, we have:*

$$D_{KL}(\mathbb{P}_\nu \| \mathbb{P}'_\nu) = \sum_{i=1}^{K} \mathbb{E}_\nu[T_i(n)] D_{KL}(P_i \| P'_i) \ .$$

Proof. We consider two instances, $\nu, \nu' \mathcal{E}_K$, with arm measures P_i and P'_i, and assume $P_i \ll P'_i$. We devise the measure $\lambda := \sum_{i \in \mathcal{K}} P_i + P'_i$. This means we can define Radon-Nikodym derivatives of all arms with respect to λ, denoted p_i and p'_i for arms P_i and P'_i respectively. Recall, that by definition:

$$D_{KL}(\mathbb{P}_\nu \| \mathbb{P}_{\nu'}) := \int \ln\left(\frac{\mathbb{P}_\nu}{\mathbb{P}_{\nu'}}\right) \mathrm{d}\mathbb{P}_\nu = \mathbb{E}_\nu\left[\ln\left(\frac{\mathbb{P}_\nu}{\mathbb{P}_{\nu'}}\right)\right] .$$

Using Lemma 2.2.4, followed by the chain rule, noting that all policy, λ, and ρ terms cancel:

$$\ln\left(\frac{\mathbb{P}_\nu}{\mathbb{P}_{\nu'}}\right) = \sum_{t=1}^{n} \ln\left(\frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)}\right) .$$

Now we apply the tower property and the fact that $p_{a_t} \mathrm{d}\lambda = \mathrm{d}P_{a_t}$ to rewrite the divergence as:

$$\mathbb{E}_\nu\left[\ln\left(\frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)}\right)\right] = \mathbb{E}_\nu\left[D_{KL}(P_{a_t} \| P_{a_t})\right] .$$

Combining both, we have:

$$\begin{aligned} D_{KL}(\mathbb{P}_\nu \| \mathbb{P}_{\nu'}) &= \sum_{t=1}^{n} \mathbb{E}_\nu\left[D_{KL}(P_{A_t} \| P_{A_t})\right] \\ &= \sum_{i \in \mathcal{K}} \mathbb{E}_\nu\left[\sum_{t=1}^{n} D_{KL}(P_{A_t} \| P_{A_t} -) \mathbb{I}\{ A_t = i\}\right] \\ &= \sum_{i \in \mathcal{K}} \mathbb{E}_\nu[T_i(n)] D_{KL}(P_i \| P'_i) . \end{aligned}$$

□

Lemma 2.2.4 is not used directly in Theorem 2.2.6, but rather in the proof of Lemma 2.2.5, which in turn is used in the proof of the theorem. This semantic aside is an artefact of the organisation of the proof and of the lemmas, but both will be required for our proof of the next theorem, which gives us a bound for the instance-independent minimax-regret.

Theorem 2.2.6 (Minimax Regret Bound, [20]). *There is $C > 0$ such that the minimax pseudo-regret of the class $\mathcal{E}_K$, when $n \geq K$, is:*

$$\bar{R}_n^*(\mathcal{E}_K) \geq C\sqrt{n(K-1)}\,.$$

Proof. Let $\mathcal{G}_K$ denote the class of bandits within $\mathcal{E}_K$ where specifically the distributions of the arms are exactly gaussian with unit variance. We will prove this result on $\mathcal{G}_K$, with $C = 1/27$, to simplify the proof for the same reasons as in the previous section.

Throughout, let π be a given policy. The idea behind the proof is to make π, which performs well on one bandit fail on another by designing the second environment entirely to trap the agent. We begin with the first environment, let $0 \leq r \leq \frac{1}{2}$, and let arm 1 have mean r and all other amrs have mean 0. In this environment we denote $\mathbb{P}_1$ the distribution on the canonical bandit $((\mathcal{K} \times \mathbb{R})^n, \mathcal{H}_n)$.

For the second environment we take the optimal arm i to be arm least taken by π in 1: $i = \operatorname{argmin}_{j \neq 1} \mathbb{E}_1[T_j(n)]$. We can now engineer a loss on this instance by taking the mean rewards from the first case and changing the reward for arm i to $2r$. It remains now to combine these to show a bound. First we will combine the regrets $\bar{R}_n^1$ and $\bar{R}_n^2$, then apply the Pinsker-type inequality (Theorem 1.2.1) and finally Lemma 2.2.5. We can bound both regrets by conditioning on $T_1(n)$ for example:

$$\bar{R}_n^1 \geq \mathbb{E}\left[\bar{R}_n^1 \middle| T_1(n) \leq \frac{n}{2}\right] P\left(T_1(n) \leq \frac{n}{2}\right) \geq \frac{nr}{2} P\left(T_1(n) \leq \frac{n}{2}\right).$$

For $\bar{R}_n^2$, do likewise but with $T_1(n) > n/2$, combining gives:

$$\begin{aligned} \bar{R}_n^1 + \bar{R}_n^2 &> \frac{nr}{2}\left(\mathbb{P}_1(T_1(n) \leq \frac{n}{2}) + \mathbb{P}_2(T_1(n) > \frac{n}{2})\right) \\ &> \frac{nr}{4} \exp\left(-D_{KL}(\mathbb{P}_1 \| \mathbb{P}_2)\right). \end{aligned} \tag{2.2}$$

Now, we use Lemma 2.2.5 to bound $D_{KL}(\mathbb{P}_1 \| \mathbb{P}_2)$.

$$D_{KL}(\mathbb{P}_1 \| \mathbb{P}_2) = 2r^2 \mathbb{E}_1[T_i(n)]\,.$$

Here we use the fact that $\sum_{j>1} \mathbb{E}_1[T_j(n)] \leq n \Rightarrow \mathbb{E}_1[T_i(n)] \leq \frac{n}{K-1}$ to further complete our bound, which upon replacement in 2.2 yields:

$$\bar{R}_n^1 + \bar{R}_n^2 \geq \frac{nr}{4} \exp\left(-\frac{2nr^2}{K-1}\right).$$

Finally, to obtain the result we take $r = \sqrt{\frac{K-1}{4n}}$, which means $n > K-1$ as $r < \frac{1}{2}$.

$$\begin{aligned}\bar{R}_n^1 + \bar{R}_n^2 &\geq \frac{1}{8}\sqrt{n(K-1)}\exp\left(-\frac{1}{2}\right) \\ &\geq \sqrt{n(K-1)} \times \frac{1}{8} \times \frac{29}{48} \\ &\geq \frac{29}{384}\sqrt{n(K-1)}\,.\end{aligned}$$

The second line follows from the power series expansion of $\exp(-\frac{1}{2})$ up to the fourth term. Finally, from $\bar{R}_n^* = \max\{\bar{R}_n^1, \bar{R}_n^2\} \geq \frac{1}{2}(\bar{R}_n^1 + \bar{R}_n^2)$, we recover:

$$\bar{R}_n^* \geq \frac{29}{768}\sqrt{n(K-1)}\,.$$

The largest simple fraction less than this is $\frac{1}{27}$. This change only tidies up the formula. □

We have thus proven that the minimax is $\Omega(\sqrt{nK})$, but the reader might wonder if this is a tight bound, and if so why it hasn't been proven. Both this result and the one in section 2.2.1 are in fact tight as there are known policies which achieve these bounds. These policies will be mentioned in the next chapter where we analyse algorithms.

Note that this theorem gives us guarantees on the worst-case regret of any policy over any possible stochastic bandit environment. In many instances the achievable regret will be less than this bound, but unavoidably, in some instances all policies will suffer regret of $\Omega(\sqrt{nK})$. This does not mean however that we should aim for an algorithm which achieves $\mathcal{O}(\sqrt{nK})$ everywhere. Indeed such an algorithm might perform well on hard instances, but will perform terribly on nicer instances where logarithmic asymptotic growth may be reached before n. There is a design trade-off between approaching the optimal minimax regret in the worst-case and maintaining optimal regret in easier instances. We will discuss this in more detail when we review algorithms in the next chapter and we will then nuance this apparent trade-off.

Chapter 3

Algorithms

BANDIT THEORY has two main components, the derivation of general results characterising specific problems, which we have covered in the previous chapter, and the design and evaluation of algorithms. Naturally, this second component will be covered in this chapter, focusing on the two main families of algorithms used in the stochastic bandit problem. First we will introduce a class of simple, intuitive, but unfortunately flawed strategies. This will serve to demonstrate to the reader that the stochastic bandit problem is not trivially solved, and that the family of more complex strategies presented in section 3.1 is noteworthy. We will give multiple regret properties, some will be proven, and we will compare them to each other and to Theorem 2.2.1.

3.1 Explore-Then-Commit

AN intuitive solution to the bandit problem would be to first deal with exploration to determine the best arm, and then move on to exploitation for the rest of the given time. This class of policies are called *Explore-Then-Commit (ETC)* strategies, a term owed to Perchet et al.[26]. They are grouped into a class as one can use any valid stopping time to determine when to stop exploration without changing some fundamental regret properties which we will present. In this section we will specifically analyse the sub-class of *fixed design* strategies, where the stopping time is a natural number, and not a random variable.

3.1.1 Algorithm

In practice this fixed design means we will explore each arm m times, to ensure uniform exploration, for a stopping time of $M := mK$. Another strategy one could consider is to explore until the mean of each arm is known with sufficient confidence relative to some parameter. While this second strategy may seem more

effective, and it indeed can be, both strategies will suffer from the same fundamental issues preventing them from achieving optimality. Therefore, we restrict ourselves to presenting the simplest strategy, the fixed design, whose algorithm can be found in Algorithm 1. For the reader's comprehension, we denote $a[\text{mod } b]$ the remainder in Euclidian division of a by b, and $\hat{\mu}_i(t)$ the sample mean of arm i at time t.

Input: m
while $t \leq N$ **do**
  if $t \leq mK$ **then**
    $A_t \leftarrow t\,[\text{mod } K] + 1$;
  end
  else
    $A_t \leftarrow \text{argmax}_{i \in \mathcal{K}} \{\hat{\mu}_i(mK)\}$;
  end
  Take action A_t;
  Store reward X_t;
end

Algorithm 1: Pseudo-code for a fixed design algorithm.

As stated, Algorithm 1, explores uniformly for M rounds, then commits to the arm with highest empirical mean. The strategy introduced, we will now make use of the results in Chapter 2 to analyse the regret of this algorithm. We will see how, and why, this strategy fails and this will highlight why we need a more flexible strategy.

3.1.2 Regret Analysis

The pseudo-regret is referred to simply as regret in this section, but the reader should recall the differences outlined in Subsection 2.1.2. First, we shall present a general formula for the regret of the ETC strategy. After discussion of this result we will use a simple case as an example, deriving an upper bound for the regret of ETC in a two-armed stochastic bandit.

Theorem 3.1.1 (General ETC regret[18]). *In a stochastic bandit setting with* 1*-subgaussian noise the pseudo-regret of the ETC algorithm satisfies for* $n \geq M$*:*

$$\bar{R}_n \leq m \sum_{i \in \mathcal{K}} \Delta_i + (n - mK) \sum_{i \in \mathcal{K}} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4} \right) .$$

Proof. We begin with the regret decomposition identity, and separate the two phases

of the algorithm, denoting i' the arm committed to and i^* the optimal arm:

$$\begin{aligned}
\bar{R}_n &= \sum_{i \in \mathcal{K}} \Delta_i \mathbb{E}[T_i(n)] \\
&= \sum_{t=1}^{m} \sum_{i \in \mathcal{K}} \Delta_i + \sum_{t=M}^{n} \sum_{i \in \mathcal{K}} \Delta_i \mathbb{E}[\mathbb{I}\{i = i'\}] \\
&= m \sum_{i \in \mathcal{K}} \Delta_i + (n - M) \sum_{i \in \mathcal{K}} \Delta_i P(i = i') \\
&= m \sum_{i \in \mathcal{K}} \Delta_i + (n - M) \sum_{i \in \mathcal{K}} \Delta_i P(\hat{\mu}_i(M) = \max_{j \in \mathcal{K}} \hat{\mu}_j(M)) \\
&\leq m \sum_{i \in \mathcal{K}} \Delta_i + (n - M) \sum_{i \in \mathcal{K}} \Delta_i P(\hat{\mu}_i(M) - \hat{\mu}^*(M) \geq 0) \\
&\leq m \sum_{i \in \mathcal{K}} \Delta_i + (n - M) \sum_{i \in \mathcal{K}} \Delta_i P(\hat{\mu}_i(M) - \hat{\mu}^*(M) - \Delta_i \geq \Delta_i)
\end{aligned}$$

See that $\hat{\mu}_i(M) - \mu_i - \hat{\mu}^*(M) + \mu^*$ is $\frac{2}{m}$-sub-gaussian, as the difference of two $\frac{1}{m}$-sub-gaussian variables, which allows us to apply Hoeffding's bound from Corollary 2.1.5, completing the proof. □

This bound is not easily interpretable, and it can't be directly compared to other algorithms due to its dependence on m. We can however, outline within it the problems ETC suffers from. In the exploration phase, the agent has incurred regret linear in m, but which decreases with smaller sub-optimality gaps. In the exploitation phase, the agent chooses the right arm with a probability which decreases with smaller gaps, and increases with greater m. This is a good illustration of the exploration-exploitation trade-off. More exploration leads to less regret by increasing confidence, but also incurs a inevitable penalty by taking sub-optimal arms. The parameter that controls this trade-off in ETC is m, but to choose a good m we need knowledge of, at least, N and preferably of the Δ_i. While the horizon may be, the sub-optimality gaps are scarcely known in practice. Knowing the gaps removes the need for a bandit altogether. To better illustrate the regret, we will focus on the simplest bandit case in Corollary 3.1.2.

Corollary 3.1.2. *[ETC regret for two-armed bandits] In the case of a two-armed Bernoulli stochastic bandit, the ETC algorithm's regret growth satisfies:*

$$\limsup_{n \to \infty} \frac{\bar{R}_n}{\ln(n)} \leq \frac{2}{\Delta} \ .$$

Proof. We apply Theorem 3.1.1 to the two-armed bandit case:

$$\begin{aligned}\bar{R}_n &\leq \frac{m}{2}\Delta + (n-2m)\Delta \exp\left(-\frac{m\Delta^2}{4}\right) \\ &\leq \frac{m}{2}\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right).\end{aligned}$$

Now we choose m dependent on n and Δ that minimises this bound, by differentiating, and we obtain, up to the rounding of m to an integer:

$$\begin{aligned}m &= \frac{4}{\Delta^2}\ln\left(\frac{n\Delta^2}{2}\right) \\ \bar{R}_n &\leq \frac{2}{\Delta}\ln\left(\frac{n\Delta^2}{2}\right) + \frac{2}{\Delta} \\ &\leq \frac{2}{\Delta}\left(1 + \ln\left(\frac{n\Delta^2}{2}\right)\right).\end{aligned}$$

Dividing by $\ln(n)$ and taking the limit superior trivially gives the result. □

Recalling from Theorem 2.2.1, and Corollary 2.2.3 that the optimal asymptotic regret is $\ln(n)/(2\Delta)$, how do ETC strategies perform? Corollary 3.1.2 suggests we are still quite far from optimality, but this could be due to our analysis. Unfortunately not. Garivier et al.[13] showed that even policies which are optimal within the class of ETC strategies only achieve asymptotic growth of $\ln(n)/\Delta$. This is half the regret of fixed design strategies, but still two times the optimal rate. This is not catastrophic, as it is only a constant difference. However, ETC strategies suffer terribly if the sub-optimality gaps are not known as there is no good way to determine a stopping time τ. In this case they have an un-improvable regret of $\mathcal{O}(n^{2/3})$[25][1]. In comparison to the minimax regret $\Omega(\sqrt{nK})$, this is an impractical gap in the domain where $K \leq \sqrt[3]{n}$. In short, ETC is viable on paper, if the sub-optimality gaps are known, which they is not the case in practice, making it highly sub-optimal.

These notably sub-optimal results come from a simple fundamental problem with all ETC strategies, which is their separation of the exploration and exploitation into two phases. The solution to achieve optimal behaviour is to design a fully sequential strategy which constantly evaluates whether to explore or exploit at each turn. Only this behaviour will allow the best arm to always be chosen asymptotically, which is key to deriving the optimal bounds. The typical, and easiest, way to design such an algorithm is to compute each round a numerical *index* for each arm

[1] Perchet et al.[26] also credit Somerville for this result, but the author could not verify this prior claim.

which measures how valuable it is to play and then play each turn the arm with highest index.

3.2 Upper Confidence Bound

IMPROVEMENTS to the explore-then-commit method are less obvious but can be understood as a different way of approaching the uncertainty in the means of arms. While we have so far attempted to quash uncertainty by finding the best arm with high confidence, in this section we will embrace the uncertainty. Using the upper bound of a fixed level confidence interval on the mean of the arms however isn't quite sufficient. The key modification is to allow arms which have been less explored to add an exploration bonus to their bound. This way, as the algorithm progresses, it plays the arm with highest upper bound for a while, which shrinks its interval, until it is over taken by either an arm with similar sample mean, or an arm which has been under-sampled. It then will repeat this process. This bonus may seem like a burden, but in fact is what will allow us to always asymptotically choose the right arm, unlike ETC strategies.

3.2.1 Algorithm

The family of algorithms called *Upper Confidence Bound (UCB)* algorithms are all based around this optimistic principle, and an exploration bonus. Rigorously UCB algorithms are a family, beginning with the work of Lai and Robbins[17] and so named by Auer et al.[4]. The algorithm we present and refer to as *UCB* algorithm is given by Lattimore and Szepesvári[22]. The reason we choose this algorithm is that it is a good middle ground between the original and more complicated UCB algorithms[12]. All UCB algorithms share the same principle and as such we must first explore the construction of the upper confidence bound itself.

Recall Hoeffding's bound (Corollary 2.1.5), which implies that for any $\epsilon > 0$, letting $\delta := \exp\left(-n\epsilon^2/2\right)$, we have $P(\hat{\mu} \geq \sqrt{2n^{-1}\ln(\delta^{-1})}) \leq \delta$. This is a plausible interval which we can adjust using the parameter δ, which requires us to assume that the $X_i - \mu_i$ are sub-gaussian. We can now deduce the smallest plausible (relative to δ, ϵ) upper bound for $\hat{\mu}_i$ to be:

$$U_i(t) := \hat{\mu}_i(t-1) + \sqrt{\frac{2}{T_i(t-1)} \ln\left(\frac{1}{\delta}\right)}.$$

We can continue this thread and choose a convenient δ, implicitly ϵ, so that the probability of ignoring the optimal arm at time t is approximately proportional to t^{-1}. This specific choice will grant us constant instead of linear regret when accounting for accidentally disregarding the best arm. Specifically, we will take $\delta^{-1} = f(t) := 1 + t\ln^2(t)$ in this particular UCB algorithm. We are now ready to introduce the UCB algorithm, whose pseudo-code is included in Algorithm 2.

while $t \leq N$ **do**
  if $t \leq K$ **then**
    $A_t \leftarrow t$
  else
    $A_t \leftarrow \text{argmax}_{i\in\mathcal{K}} \left\{\hat{\mu}_i(t-1) + \sqrt{\frac{2\ln f(t)}{T_i(t-1)}}\right\}$
  end
  Take action A_t;
  Store reward X_t;
end

Algorithm 2: Pseudo-code for the UCB algorithm

Variations on the theme of UCB often revolve around the specific index (UCB1[4]) or the use of other mechanics (UCB2[4]). There are also more general formulations which take several arguments[7], and expand the range of confidence bounds and exploration bonuses possible. We will analyse the regret only of this formulation of UCB, but more general results exist.

3.2.2 Regret Analysis

Using the tools developed in section 2.1 we will now prove several results about the regret of the UCB algorithm. Using the results from sections 3.1 and 2.2, we will be able to compare the results to other algorithms and optimal policies. To begin, instance-dependent bounds on the finite-time and asymptotic regret will be given in 3.2.1. The finite-time bound serves to demonstrate the methodology of regret bound proofs, and will allow us to easily derive the asymptotic bounds which by Corollary 2.2.2 shows this UCB algorithm to be asymptotically optimal for gaussian bandits.

Theorem 3.2.1 (UCB Regret Bounds, [22])**.** *The pseudo-regret of the UCB algorithm in a stochastic bandit satisfies:*

1. $\bar{R}_n \leq \sum_{i:\Delta_i>0} \inf_{\epsilon\in(0,\Delta_i)} \left\{1 + \frac{5}{\epsilon} + \frac{2}{(\Delta_i-\epsilon)^2}\left(1 + \ln(f(n)) + \sqrt{\pi\ln(f(n))}\right)\right\}$,

2. $\limsup_{n\to\infty} \frac{\bar{R}_n}{\ln(n)} \leq \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}$.

Before proving this theorem, we present and prove a lemma which will be required during the proof. This lemma provides a bound on the expectation of the sum of indicator variables of the form *a confidence interval's upper bound is greater than a value*.

Lemma 3.2.2 ([22]). *Let $X_i - \mu$ be IID sub-gaussian random variables, take $\epsilon > 0$ and let:*

$$\hat{\mu}_t := \frac{1}{t}\sum_{i=1}^{t} X_i,\ \kappa = \sum_{t=1}^{n} \mathbb{I}\left\{\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \epsilon\right\}.$$

Then $\mathbb{E}[\kappa] \leq 1 + \frac{2(a+\sqrt{a\pi}+1)}{\epsilon^2}$.

Proof. Let $u = 2a\epsilon^{-2}$, starting with the definition of κ we have:

$$\begin{aligned}
\mathbb{E}[\kappa] &= \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{I}\left\{\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \epsilon\right\}\right] \\
&= \sum_{t=1}^{n} P\left(\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \epsilon\right) \\
&\leq u + \sum_{t=\lceil u\rceil}^{n} P\left(\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \epsilon\right).
\end{aligned}$$

From Theorem 2.1.4, it follows that:

$$\begin{aligned}
\mathbb{E}[\kappa] &\leq u + \sum_{t=\lceil u\rceil}^{n} \exp\left(-\frac{t}{2}\left(\epsilon - \sqrt{\frac{2a}{\epsilon}}\right)^2\right) \\
&\leq 1 + u + \int_{u}^{\infty} \exp\left(-\frac{t}{2}\left(\epsilon - \sqrt{\frac{2a}{\epsilon}}\right)^2\right) \mathrm{d}t \\
&\leq 1 + \frac{2a}{\epsilon^2} + \frac{2(\sqrt{a\pi}+1)}{\epsilon^2}.
\end{aligned}$$

□

Proof. 1. This proof is based upon the regret decomposition identity (Theorem 2.1.2), where we will bound $\mathbb{E}[T_k(n)]$. We will investigate separately the two

possible scenarios leading to playing the suboptimal arm. First it is possible that our upper bound $U_i(t)$ is under the true value of $\mu_i(t) - \epsilon$: we have vastly underestimated the payoff of the optimal arm. In the second possible case there is a suboptimal arm whose exploration penalty leads it to be chosen over the optimal arm. Formally we will separate $T_i(n) = \sum_{t=1}^{n} \mathbb{I}\{A_t = i\}$ into S_1 and S_2 corresponding to each case. Let μ_o denote the mean of the optimal arm.

$$S_1 = \sum_{t=1}^{n} \mathbb{I}\left\{\hat{\mu}_o(t-1) + \sqrt{\frac{2\ln f(t)}{T_i(t-1)}} \leq \mu_o - \epsilon\right\}.$$

$$\mathbb{E}[S_1] = \sum_{t=1}^{n} P\left(\hat{\mu}_o(t-1) + \sqrt{\frac{2\ln f(t)}{T_i(t-1)}} \leq \mu_o - \epsilon\right)$$

Now redefine the rewards in terms of s the number of times a specific arm is pulled instead of t. Let $(Z_{i,s})_s$ be a sequence of iid rewards from arm s. Note that $X_t = Z_{A_t,T_{A_t}(t)}$, and let $\hat{\mu}_{i,s} = \frac{1}{s}\sum_{j=1}^{s} Z_{i,j}$. This replaces the random $T_i(n)$ with a constant s, and summing over the range of $T_i(n)$ we obtain:

$$\mathbb{E}[S_1] \leq \sum_{t=1}^{n}\sum_{s=1}^{n} P\left(\hat{\mu}_{o,s} + \sqrt{\frac{2\ln f(t)}{s}} \leq \mu_o - \epsilon\right).$$

Applying Theorem 2.1.4, then simplifying with common identities and recalling $f(t) = 1 + t\ln^2(t)$, so that $\sum_{t=1}^{\infty}\frac{1}{f(t)} < \frac{5}{2}$ by numerical evaluation:

$$\begin{aligned}
\mathbb{E}[S_1] &\leq \sum_{t=1}^{n}\sum_{s=1}^{n} \exp\left(-\frac{s}{2}\left(\sqrt{\frac{2\ln f(t)}{s}} + \epsilon\right)^2\right) \\
&\leq \sum_{t=1}^{n}\frac{1}{f(t)} \times \sum_{s=1}^{n}\exp\left(\frac{-s\epsilon^2}{2}\right) \\
&\leq \sum_{t=1}^{\infty}\frac{1}{f(t)} \times \sum_{s=1}^{n}\exp\left(\frac{-s\epsilon^2}{2}\right) \\
&\leq \frac{5}{2} \times \int_0^{\infty}\exp\left(\frac{s\epsilon^2}{2}\right)\mathrm{d}s \\
&\leq \frac{5}{\epsilon^2}.
\end{aligned}$$

Now, rearranging S_2 into a form to which we can apply Lemma 3.2.2 yields:

$$\begin{aligned} S_2 &= \sum_{t=1}^{n} \mathbb{I}\left\{\hat{\mu}_i(t-1) + \sqrt{\frac{2\ln f(t)}{T_i(t-1)}} \geq \mu_o - \epsilon, A_t = i\right\} \\ \mathbb{E}[S_2] &\leq \mathbb{E}\left[\sum_{s=1}^{n} \mathbb{I}\left\{\hat{\mu}_{i,s} + \sqrt{\frac{2\ln f(t)}{s}} \geq \mu_o - \epsilon\right\}\right] \\ &= \mathbb{E}\left[\sum_{s=1}^{n} \mathbb{I}\left\{\hat{\mu}_{i,s} - \mu_i + \sqrt{\frac{2\ln f(t)}{s}} \geq \Delta_i - \epsilon\right\}\right] \\ &\leq 1 + \frac{2}{(\Delta_i - \epsilon)^2}\left(\ln f(n) + \sqrt{\pi \ln f(n)} + 1\right). \end{aligned}$$

Combining S_1 and S_2, and completing the regret decomposition identity, leads to the desired result. The infimum ensures this bound is minimised for ϵ, while not allowing the denominators in the regret bound to be zero.

2. Taking $\epsilon = \ln^{-1/4}(n)$ in the first part of the theorem gives:

$$\begin{aligned} \bar{R}_n &\leq \sum_{i:\Delta_i>0} \inf_{\epsilon\in(0,\Delta_i)} \left\{1 + 5\ln^{\frac{1}{4}} + \frac{2}{\left(\Delta_i - \ln^{-\frac{1}{4}}(n)\right)^2}\left(1 + \ln(f(n)) + \sqrt{\pi\ln(f(n))}\right)\right\} \\ &\leq \sum_{i:\Delta_i>0} 1 + \inf\left\{5\sqrt[4]{\ln(n)} + \frac{2\sqrt{\ln(n)}}{\left(\Delta_i\sqrt[4]{\ln(n)} - 1\right)^2}\left(1 + \ln(f(n)) + \sqrt{\pi\ln(f(n))}\right)\right\} \\ &\leq \sum_{i:\Delta_i>0} 1 + \inf\left\{5\sqrt[4]{\ln(n)} + \frac{2\sqrt{\ln(n)}}{\Delta_i\sqrt{\ln(n)}}\left(1 + \ln(f(n)) + \sqrt{\pi\ln(f(n))}\right)\right\} \end{aligned}$$

Substituting in $f(n)$ and dividing by $\ln(n)$ and taking the limit superior yields the result. □

The bonds of Theorem 3.2.1 can be simplified for legibility trivially by a clever choice of ϵ, which is given in Corollary 3.2.3.

Corollary 3.2.3 (Simplified UCB Regret Bounds, [22]). *Choosing $\epsilon = \frac{\Delta_i}{2}$ in Theorem 3.2.1 gives:*

$$\bar{R}_n \leq \sum_{i:\Delta_i>0}\left[\Delta_i + \frac{8}{\Delta_i}\left(\ln f(n) + \sqrt{\pi\ln f(n)} + \frac{7}{2}\right)\right].$$

Furthermore, for all $n \geq 2$, there is some strictly positive universal constant C such that:

$$\bar{R}_n \leq \sum_{i:\Delta_i>0} \left(\Delta_i + \frac{C \ln(n)}{\Delta_i} \right) .$$

From Theorem 3.2.1 and Corollary 2.2.2, as stated, we can see that this UCB algorithm is asymptotically optimal for the class of gaussian bandits. There are related algorithms in the UCB family, such as KL-UCB[12] which is optimal in the case of Bernoulli bandits, and its variants, which specialise in achieving the regret bound of Theorem 2.2.1 for various specific classes of stochastic bandits. Thus UCB policies are asymptotically optimal. This result also demonstrates the improvement compared to ETC strategies described earlier. Finally, we conclude this analysis by deriving an instance dependent bound, which we will be used to compare the UCB algorithm to the minimax regret achievable by Theorem 2.2.6.

Theorem 3.2.4 (Order of UCB Regret, [19]). *The pseudo-regret of a worst-case instance of the UCB algorithm with Δ_i not small for all i satisfies:*

$$\bar{R}_n = \mathcal{O}\left(\sqrt{Kn \ln(n)} \right) .$$

Proof. Fixing $\Delta > 0$, we have $\mathbb{E}[T_i(n)] \leq C \ln(n) \Delta_i^{-1}$ from which we obtain a distribution free bound:

$$\begin{aligned}
\bar{R}_n &= \sum_{i=1}^{n} \Delta_i \mathbb{E}[T_i(n)] \\
&\leq \sum_{i:\Delta_i<\Delta} \Delta_i \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i \geq \Delta} \Delta_i \mathbb{E}[T_i(n)] \\
&\leq n\Delta + \sum_{i:\Delta_i \geq \Delta} \frac{C \ln(n)}{\Delta_i} \\
&\leq n\Delta + K \frac{C \ln(n)}{\Delta} .
\end{aligned}$$

Minimising the bound by letting $\Delta = \sqrt{K \ln(n) n^{-1}}$ gives the result. □

This is only an $\sqrt{\ln(n)}$ factor away from being minimax optimal, which justifies calling UCB policies *near-minimax optimal*. A noteworthy alternative to UCB, while still a related index algorithm, is the MOSS algorithm by Audibert and Bubeck[3]. It achieves a different asymptotical regret, although it is of the same logarithmic order, but achieves the minimax bound of $\mathcal{O}(\sqrt{nK})$, proving that $\bar{R}_n(\mathcal{E}_K) = \Theta(\sqrt{n(K)})$.

Summary

FOR the reader's convenience, we collate here the most important results and comments made in this thesis. This is to allow for easier comparison and comprehension of how these pieces fit together. However, the keys to thorough understanding can not all be collected here, in particular results from the first chapter will not be covered. Mirroring the structure of the thesis we will first give the characterisation of the problem then compare algorithms.

First, let us recall what the key assumptions about $\mathcal{E}$ are. We assumed that rewards had low-variance, by either being subgaussian, but most importantly we assumed the environment is stationary and rewards are independent within and between arms. We defined the pseudo-regret $\bar{R}_n$, to be the loss relative to perfect information, and chose it as the measure of an agent's performance. Under the above assumptions, in the stochastic bandit problem the optimal asymptotical regret growth attainable was shown by Lai and Robbins[17] to be:

$$\bar{R}_n \sim \sum_{i \neq i^*} \frac{\Delta_i}{D_{KL}(P_i \| P_{i^*})} \ln(n) \,.$$

This bound helps understand where difficulty lies in a bandit instance. Large gaps aren't so perilous if the underlying distributions are easily distinguished. Conversely, near-identical distributions with very small differences won't lead to a high regret as the large number of wrong guesses is compensated by the low cost of each error. This is asymptotic however and in on a hard instance the logarithmic domain may never be reached. In this case, we showed the minimax, the best achievable regret on the worst instance on $\mathcal{E}_K$, to be:

$$\bar{R}^*_n(\mathcal{E}_K) = \Theta\left(\sqrt{n(K-1)}\right) \,.$$

These two bounds allow us to evaluate algorithms based on their flexibility to difficult instances, with the minimax, and on their probable long-run behaviour on nicer instances, with the asymptotic bound. One would like to balance behaviour in both these domains, as a policy that for example achieves $\mathcal{O}(\sqrt{nK})$ everywhere is minimax-optimal but vastly sub-optimal in all other instances. To the knowledge of the author there is no known policy which is both asymptotically and minimax

optimal.

Recall that all strategies that explore until a certain stopping time and later exploit the best arm were grouped in the class of ETC strategies. While they are intuitive, ETC strategies are held back by the ability to determine a good stopping time, which requires they know the horizon N and if possible the sub-optimality gaps Δ_i. In practice, the gaps are almost never known, and without them ETC strategies incur a regret of $\Omega(n^{2/3})$. In comparison to the minimax, this is highly sub-optimal. ETC strategies also fail on nicer instances as their regret growth can not be less than twice the optimal rate. Beyond the stopping time issue, the reason for the necessary sub-optimality of ETC is that the fundamental exploration-exploitation trade-off must be addressed at each time-step. Exploration and exploitation *must* both be spread out throughout the whole run-time.

Considering fully sequential strategies, we showed that the UCB family of algorithms can achieve the Lai and Robbins[17] bound in asymptotic regret. Using confidence intervals designed with Hoeffding's bound we can naturally incorporate into an index a bonus for under-explored arms while still playing optimistically the arm of highest return most of the time. Members of this family achieve optimality in asymptotical regret growth, and are near-minimax optimal with a minimax regret of $\mathcal{O}(\sqrt{nK\ln(n)})$. We have reviewed one member[22] and mentioned several more[3,4,12,17] in this thesis. Each variation on the theme of an optimistic index algorithm share these properties to some extent. For example MOSS[3] achieves the minimax regret order, but its asymptotic growth is $\mathcal{O}(K\ln(n))$, which is sub-optimal. The exact tuning of the algorithm's index guarantees its effectiveness in some settings at the price of sub-optimality in others. There is an implementation problem, then, of choosing exactly which UCB algorithm to use. This helps explain the range of algorithms in the family.

The author hopes that the reader has come away from this thesis comfortable enough with the minimax concept to take note of an essential fact. The minimax regret is exactly the regret order one can achieve in the adversarial bandit. If the reader recalls our proof, we designed an environment to trick our policies. We were, in fact, taking on the role of the adversary and bending the randomness of rewards to fit our design and trick the agent. It is no surprise then that this is also the worst regret in the adversarial case. This is a good reason in practice to not worry too much about exact minimax optimality in stochastic bandits, since if one is trying to compete in a very difficult environment reframing it as an adversarial bandit and applying optimal adversarial algorithms is more sensible. As this exemplifies, the best way to improve overall regret is to tailor the algorithm to the problem to exploit any and all available information by changing the paradigms and setting.

Index of Notation

This index orders by domain the symbols used in this thesis which have an intrinsic meaning. Unless listed as different here, notation carries over from one chapter to the next.

Bandit Theory

$((\mathcal{K} \times \mathbb{R})^N, \mathcal{H}_N)$ The canonical bandit

$\bar{R}_n$ The pseudo-regret in round n

$\bar{R}_n^*(\cdot)$ The minimax pseudo-regret of the class of bandits

Δ_i The sub-optimality gap for arm i: $\mu^* - \mu_i$

$[\![\cdot]\!]$ The interval in $\mathbb{N}$ from 1 up to $\cdot$

$\mathbb{I}$ The indicator function

$\mathcal{E}$ The class of stochastic bandit instances

$\mathcal{E}_K$ The class of all K-armed stochastic bandit instances

$\mathcal{G}_K$ The class of K-armed stochastic gaussian bandits

$\mathcal{K}$ The arm set

$\mathcal{O}(\cdot)$ "Big O" complexity theory notation

μ^* The mean of the optimal arm

ν, ν' Stochastic bandit instances

$\Omega(\cdot)$ "Big Omega" complexity theory notation

$\perp$ Independence

π A policy

$\mathbb{P}_\nu$ The canonical measure on instance ν

$\Theta(\cdot)$ "Big Theta" complexity theory notation

A_t, I_t The action at time t, i.e. the arm played

$f \sim \cdot$ f grows asymptotically as

H_n The History

K The number of arms

M, m Quantities related to fixed design strategies

N The horizon, or number of rounds

P_i (or P_i^ν) The Probability measure of arm i, equivalently, its distributions (in instance ν)

R_n The regret in round n

$T_i(n)$ The number of times arm i is played up to round n

$U_i(t)$ The upper confidence bound on arm i at round t

$X_{i,t}$ The reward observed by playing arm i at time t

Information Theory

λ, μ Probability measures

$D_{KL}(\cdot\|\cdot)$ The Kullback-Leibler Divergence

Measure Theory

$(\mathcal{X}, \mathcal{S})$ A measure space

$(\mathcal{X}, \mathcal{S}, \mu)$ A measurable space

$\cdot \ll \cdot$ Absolute continuity

$\int$ The Lebesgue integral

$\mathcal{B}$ The Borel σ-algebra

$\mathcal{B}^*$ The extended Borel σ-algebra

$\mathcal{E}$ A class of sets (e.g. a σ-algebra)

$\mathcal{L}(\cdot)$ The Lebesgue σ-algebra over the set

$\mathcal{P}\cdot$ The power set

$\mathcal{S}, \mathcal{F}$ σ-algebras

$\mathcal{X}$ A topological space

$\mu(\cdot)$ A measure

Ω A sample space

$\mathbb{R}^*$ The extended real line $\mathbb{R} \cup \{-\infty, \infty\}$

$\sigma(\cdot)$ The σ-algebra generated by the set

$f(\cdot)$ A density, defined as a Radon-Nikodym derivative

Miscelaneous

: "such that"

$\emptyset$ The empty set

Bibliography

[1] Agrawal, S. and N. Goyal
2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, S. Mannor, N. Srebro, and R. C. Williamson, eds., volume 23 of *Proceedings of Machine Learning Research*, Pp. 39.1–39.26, Edinburgh, Scotland. PMLR.

[2] Ash, R. B.
1965. *Information Theory*. New York: Wiley.

[3] Audibert, J.-Y. and S. Bubeck
2009. Minimax policies for adversarial and stochastic bandits. In *COLT*, Pp. 217–226.

[4] Auer, P., N. Cesa-Bianchi, and P. Fischer
2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

[5] Auer, P., N. Cesa-Bianchi, Y. Freund, and R. E. Schapire
1995. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, Pp. 322–331. IEEE.

[6] Awerbuch, B. and R. D. Kleinberg
2004. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, Pp. 45–53. ACM.

[7] Bubeck, S., N. Cesa-Bianchi, et al.
2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

[8] Burnetas, A. N. and M. N. Katehakis
1996. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142.

[9] Cover, T. M. and J. A. Thomas
2012. *Elements of information theory*. John Wiley & Sons.

[10] Dudík, M., D. J. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang
2011. Efficient optimal learning for contextual bandits. *CoRR*, abs/1106.2369.

[11] Feinstein, A.
1958. *Foundations of information theory*. McGraw-Hill.

[12] Garivier, A. and O. Cappé
2011. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, S. M. Kakade and U. von Luxburg, eds., volume 19 of *Proceedings of Machine Learning Research*, Pp. 359–376. PMLR.

[13] Garivier, A., T. Lattimore, and E. Kaufmann
2016. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, Pp. 784–792.

[14] Garivier, A., P. Ménard, and G. Stoltz
2016. Explore first, exploit next: The true shape of regret in bandit problems. *ArXiv e-prints*.

[15] Gittins, J., K. Glazebrook, and R. Weber
2011. *Multi-armed bandit allocation indices*. John Wiley & Sons.

[16] Kullback, S.
1997. *Information theory and statistics*. Courier Corporation.

[17] Lai, T. L. and H. Robbins
1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

[18] Lattimore, T. and C. Szepesvári
2016a. First steps: Explore-then-commit. `http://banditalgs.com/2016/09/14/first-steps-explore-then-commit/`.

[19] Lattimore, T. and C. Szepesvári
2016b. Instance dependent lower bounds. `http://banditalgs.com/2016/09/30/instance-dependent-lower-bounds/`.

[20] Lattimore, T. and C. Szepesvári
2016c. More information theory and minimax lower bounds. `http://banditalgs.com/2016/09/28/more-information-theory-and-minimax-lower-bounds/`.

[21] Lattimore, T. and C. Szepesvári
2016d. Stochastic linear bandits and ucb. http://banditalgs.com/2016/10/19/stochastic-linear-bandits/.

[22] Lattimore, T. and C. Szepesvári
2016e. The upper confidence bound algorithm. http://banditalgs.com/2016/09/18/the-upper-confidence-bound-algorithm/.

[23] Leadbetter, R., S. Cambanis, and V. Pipiras
2014. *A basic course in measure and probability: Theory for applications*. Cambridge university press.

[24] Mannor, S. and O. Shamir
2011. From bandits to experts: On the value of side-observations. *CoRR*, abs/1106.2436.

[25] Maurice, R. J.
1957. A minimax procedure for choosing between two populations using sequential sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 19(2):255–261.

[26] Perchet, V., P. Rigollet, S. Chassang, E. Snowberg, et al.
2016. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681.

[27] Rivasplata, O.
2012. Subgaussian random variables: An expository note.

[28] Robbins, H.
1985. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, Pp. 169–177. Springer.

[29] Rosenblatt, F.
1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

[30] Samuel, A. L.
1959. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229.

[31] Shannon, C. E.
2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

[32] Somerville, P. N.
1954. Some problems of optimum sampling. *Biometrika*, 41(3/4):420–429.

[33] Sutton, R. S. and A. G. Barto
1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

[34] Thompson, W. R.
1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

[35] Valko, M., R. Munos, B. Kveton, and T. Kocák
2014. Spectral bandits for smooth graph functions. In *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing and T. Jebara, eds., volume 32 of *Proceedings of Machine Learning Research*, Pp. 46–54, Bejing, China. PMLR.

[36] Vaswani, S. and L. V. S. Lakshmanan
2015. Influence maximization with bandits. *CoRR*, abs/1503.00024.

[37] Vernade, C., O. Cappé, and V. Perchet
2017. Stochastic bandit models for delayed conversions. *CoRR*, abs/1706.09186.

[38] Yue, Y. and T. Joachims
2009. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, Pp. 1201–1208. ACM.