Mathematics & Statistics | Lancaster University

MULTI-ARMED BANDITS

Sequential Decision Agents in Stochastic Environments

by

Lorenzo Croissant

Supervisor:

Dr. Azadeh Khaleghi

Dissertation submitted 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, wether 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 later's LaTeX dissertation template[1] provided the ground work for the typography of this thesis. As both are public templates, this thesis' template 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 Pr. 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 it set out with different goals, the contents of this thesis 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

How should one allocate finite resources amongst multiple different projects whose progress is unpredictable in order to achieve the greatest success rate? What is the minimal amount of samples needed to identify the distribution of highest mean amongst a set of unknown distributions? How often should a recommender system suggest completely new objects in order to map out your preferences? Suppose the recommender only knows wether 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? How much more difficult is an allocation problem if the environment is markovian instead of stationary? All these problems can be formulated and answered in the machine learning framework of bandit theory.

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 what-so-ever 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* receives *bandit feedback*, a small amount of information that depends 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 for the specific problem considered

There are a very wide range of problems in bandit theory, which fall into three natures. From the most general to the most specific, we first have *adversarial bandits*, where no assumptions are made about the slot machines, there reward is set at each round by an opponent which uses a possibly randomised procedure. Second, in *Markovian bandits*, each slot machine is a unknown Markov machine, whose state may change at each round. The final case is the special case of *stochastic bandits*, where each Markov machine is stationary. As is generally the case in mathematics, the specific case is a simpler problem than the more general formulation. Due to space limitations, this thesis will only provide an introduction to stochastic bandits, as the author feels that any survey of Markovian or adversarial bandits must be supported by a solid foundation in stochastic bandits.

In parallel, there are many settings one can consider in at least one of these three natures which give bandit theory a lot of flexibility to specialise to specific real-world problems. For example, in many real-world applications of information systems, the agent's decisions can be grounded in some external data, called a *context*. The family of bandit problems with contexts are thus called *contextual bandits*, and have been the

subject of a lot of research[26]. There are many other examples, which the reader may find of interest, such as: bandits with side observation[29], bandits on combinatorial structures[18], duelling bandits[43], bandits with delayed feedback[10]. We will not cover any specialised setting and simply present the regular stochastic bandit for the same reasons as our restriction to the stochastic case.

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 $t \in \mathbb{N}$ exactly one action in a set $\mathcal{K}$ of size K, immediately receives a reward X_t for the arm it played and then moves to round $t+1$. 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].

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. 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.

Before we begin the main discussion of this thesis, the author would like to invite the reader to pause with him and explore the history of the field of bandit theory. While inherently interesting, at least to the author, the reason to do so is to understand how bandit theory fits into the wide picture of machine learning, sampling and experimental design, project management, and control theory, to cite only a few. This review of the literature intends to highlight connections to other fields, evolutions of methods, and provide some answers to the questions which opened this text.

The first work in bandit theory is considered to be Thompson[39], which involved the design of sequential clinical trials for two drugs with the goal of designing strategies to minimise the number of people given the least effective drug. Taking as reward the survival or death of each patient this is a stochastic 2-armed bandit with Bernoulli arms. While Thompson's work did not lead to an immediate flurry of works in bandit theory, his bayesian approach proved remarkably effective and is still studied today[1]. The IID stochastic bandit, which is our subject in this thesis, was introduced in Robbins[33] building on the sequential design work of Wald. The

[2]This issue will be nuanced in section 2.1.2, 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 $[0,1]$.

most important result in stochastic bandits came three decades later in Lai and Robbins[19], which gives asymptotical guarantees for the performance of any policy an agent could use. This seminal work also proposed a policy which attains this optimality, we will discuss both these results in this thesis. This proof was then extended by Burnetas and Katehakis[8] to a larger family of distributions, while the policy was named UCB and improved by Auer et al.[4]. From the 1990s onwards many settings of the bandit problem were explored.

The markovian bandit

In order to provide an exhaustive presentation of the regular stochastic bandit problem, we will begin by a quick outline of foundational material in measure theory which will allow us to redefine probability, random variables, integration and finally density. We will repeat this exercise with information theory, presenting a quantity of the difference between two distributions, and some inequalities related to it. With these foundational principles we will then 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 the next section, 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 document.

Chapter 1

Foundational Preliminaries

RIGOUROUS definitions are the paramount prerequisite to any mathematical exercise, and bandit theory will not escape this fact. While not constitutive of the main body of bandit theory, notions outlined in this chapter are required to proceed in an orderly manner in this 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. Will be presented only the required components upon which to base the rest of this thesis. We will begin with a march through measure theory, defining or redefining the concepts of probability, including integration, the density, and measures. We will then immediately apply these building blocks to a cherry-picked brief in Information theory, whose purpose is to present to the reader the Kullback-Leibler divergence and some of its properties.

1.1 Elements of Measure Theory

MEASURE THEORY allows us to reformulate and generalise many statements about probability distributions. This new field refines the theory of probabilities and allows us to derive a new understanding of what probabilities, events, outcomes, and random variables are. We will assume that the basic notions of sigma-algebras, sample space and of random variables are known in an intuitive formulation. 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.[28].

1.1.1 Measures

We consider first an arbitrary topological space $\mathcal{X}$, and we will take interest 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[28, 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 component of probability theory, the random variable. In the next subsection we will rebuild the idea of a distribution, 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. Allowing us to formulate a sensible definition for the integral of these 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. 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 ([28, p. 69]). *If f is integrable with respect to a measure μ, it is finite almost everywhere with respect to μ. Furthermore, if f is measurable, defined almost everywhere and E is a set with 0 measure, then $\int_E f \mathrm{d}\mu = 0$.*

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 in 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[28, 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 ([28, 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[28, p. 103])**.** *Let μ, ν be σ-finite measures on $(\mathcal{X}, \mathcal{S})$, and λ be a σ-finite measure on $\mathcal{S}$. 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

PART OF 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 original purpose of information theory to illustrate its fundamentals. Using the results from section

1.1, we will diverge from the original works of C.E. Shannon[36], 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[17]. This is because we do not truly need core information theory, but the parts which boil over into 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[36] 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 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. 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, adapted from Kullback[17], 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 and H_2. We can then define the mean information using the generalised probability densities f_1, f_2 for $\mu_1(E) \neq 0$:

$$I(1, 2; E) = \frac{1}{\mu_1(E)} \int_E \ln \frac{f_1(x)}{f_2(x)} \mathrm{d}\lambda_1$$

$$= \frac{1}{\mu_1(E)} \int_E f_1(x) \ln \frac{f_1(x)}{f_2(x)} \mathrm{d}\mu \,.$$

The second line follows from the Radon-Nikodym derivative $f_1(x) = \frac{\mathrm{d}\lambda_1}{\mathrm{d}\mu}(x)$. Here we notice that f is not important and instead can be fully described by two metrics, μ and λ_1. 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:

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

This quantity is called the Kullback-Leibler (KL) divergence. Note that if f_1 is not absolutely continuous with respect to f_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_1 to λ_2. 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 both measures. 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 aforementioned 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. This identity is given in theorem 1.2.1 and is proven immediately afterwards.

Theorem 1.2.1 (Pinsker-type inequality[24]). *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 one-armed bandit for instance, one gambles against a seemingly random slot machine. The slot machine however, 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, after which we will show a quintessential theorem called the regret decomposition identity. After this, 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.

2.1.1 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. Afterwards we will formulate the regret decomposition identity, which will prove to be a fundamental result. The regret is a random variable, R_n, defined as the difference between repeating the action of highest summed rewarded and the received sum of rewards[7]. In a way, 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 will tend to examine the weaker 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 terminological sloth.

Definition 2.1.1 (Pseudo-Regret of a stochastic bandit)**.** In a stochastic bandit 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.

Definition 2.1.2 (Tractable pseudo-regret of a stochastic bandit)**.**

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

Proof. Starting with definition 2.1.1 we rewrite the formula for R_n:

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

$$= \max_{i \in \mathcal{K}} \{n\mu_i\} - \sum_{t=1}^{n} \mu_{I_t}$$
$$= n\mu^* - \sum_{t=1}^{n} \mu_{I_t} .$$

□

These two definitions are all we need to introduce the main result of this section, whose proof will be immediately given thereafter.

Theorem 2.1.1 (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* $T_k(n) := \sum_{t=1}^{n} \mathbb{I}\{I_t = k\}$ *be the random variable counting the number of times arm k is chosen in n rounds. We can now decompose the pseudo-regret in terms of* Δ_k *and* $\mathbb{E}[T_k(n)]$*:*

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

Proof. Recalling from the tractable pseudo-regret, we will re-index the sums 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 2.1.1 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.

2.1.2 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: $\hat{\mu}_k := \frac{1}{n_k}\sum_{t=1}^{n} X_{I_t,t}\mathbb{I}\{I_t = k\}$. 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. error) 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 which forbids its use. Instead we will define a new property of a random variable which will allow us to derive new properties about its tail probabilities.

Definition 2.1.3 (Sub-gaussianity)**.** A random variable X is σ^2-sub-gaussian if it satisfies:

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

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 2.1.2 (Properties of sub-gaussian random variables[22])**.** *Let X be a sub-gaussian random variable.*

- *We have $\mathbb{E}(X) = 0$ and $\text{var}(X) \leq \sigma^2$.*
- *For $c \in \mathbb{R}$, cX is $c^2\sigma^2$-sub-gaussian.*
- *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.

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

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

Using the second order taylor expansion, we obtain for all $\lambda \in \mathbb{R}$:

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

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 $\frac{\lambda^2}{2} > 0$, and take the limit as $\lambda \to 0$:

$$\text{var}\,(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \leq \frac{1}{\sigma^2} .$$

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

◦ 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 2.1.3 (Concentration of Sub-gaussian Random Variables[22]). *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:

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

Now we choose λ to minimise this bound as it holds for all $\lambda > 0$. See 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 2.1.3, we can derive a bound for the sample mean we were interested in.

Corollary 2.1.4 (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

GENERAL results about the regret behaviour of algorithms in the stochastic bandit problem will allow us to establish benchmarks against which we can compare the effective performance of the algorithms which will be discussed in chapter 3. We will present the key behaviour of regret growth, characterising the types of policies which we will take as valid, and then proving a tight lower bound on the general-case asymptotic regret growth attainable, the seminal result in bandit theory. This first analysis will focus on what can be achieved on nice instances, the first benchmark, while the second will analyse the best-achievable worst case regret, a concept borrowed from game theory, which will be defined. This second benchmark allows us to compare performance of algorithms in the most gruelling environments, where the random distributions are virtually equivalent to playing against an adversary, bent on tricking the algorithm.

2.2.1 Optimal Asymptotic Regret Growth

The key result in this section will show that any strategy which is *consistent* achieves at most logarithmic asymptotic regret growth in terms of the number of rounds, up to a universal constant which will make this bound tight. It is not trivial however to see why this is, nor to understand it. It is important firstly, that this behaviour is asymptotic, and must only be considered if N is large in regards to the difficulty of the instance, and the number of arms. Since we must intuitively explore the whole action space to hope to find the optimal arm, algorithms often start by pulling each arm one or more times, denoted m. This leads to a linear regret term for the first mK rounds, which does not interfere with the result. A further problem with asymptotic bounds, is that in a finite time instance, the asymptotic domain may never be reached[14], leading to non logarithmic growth. Notwithstanding, asymptotic guarantees are still a valuable measure for algorithms, even if they must be taken with the usual caution for asymptotic results.

In deriving a general result for all acceptable policies, the next issue we are faced with is exactly how to define an acceptable policy. Of course we can eliminate trivial policies which are inapplicable in practice, such as $\pi(t) = \text{argmax}_{k \in \mathcal{K}} \mu_k$, which requires knowledge of the best arm ahead of time, while achieving 0 regret. Thus our policy π must not take into account the means μ_i, and therefore not the suboptimality gaps Δ_i either. Inversely, we are only interested in strategies which have some intelligence. We are not interested in bounding the regret growth of a policy which simply plays an arm at random each round, forever. Just as the previous class

had inapplicable knowledge, this class has inapplicable linear regret. Surprisingly to the reader, perhaps, these classes aren't so different. Note that any policy that has knowledge of an instance ν will perform linearly on an instance ν'. Take the policy, $\pi(t) = \text{argmax}_{k \in \mathcal{K}(\nu)} \mu_k(\nu)$ given above, in an instance with another optimal arm it suddenly incurs linear regret. Thus, we would want to simply restrict ourselves to strategies which are sub-linear. For historical reasons, in the literature the condition is made slightly more stringent, requiring sub-polynomial growth. This was the setting chosen by Lai and Robbins[19], and remains in practice today. Note, however, that this is not a monolithic reverence for the fathers of bandit theory, this requirement is well grounded in fact. Without elaborating on the contents of the next section, there are abstract policies which are sub-polynomial with fractional powers over all instances, and this will be proven. Therefore, we might sensibly limit ourselves to sub-polynomial policies in our analysis, as they are uniformly better than ones which are simply sub-linear.

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

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

This definition, and its justification out of the way, we may proceed to the main result of this section, theorem 2.2.1. This theorem gives an instance dependent lower bound for the asymptotic regret growth, with remarkable elegance. This result connects the intuitive difficulty of the problem, the number of arms, (the length of the sum), and the gaps between arms with the Kullback-Leibler divergence between the means of each arm. For large gaps, thus extremely different distributions, the resulting regret will be small as the arms are easily distinguished. Even for small gaps, the regret is small if arms are very different, but likely moderate if the arms are quite similar. The divergence really highlights the intrinsic difficulty of the problem.

Theorem 2.2.1 (Assymptotic Growth Lower-Bound[19, thm. 1])**.** *For all consistent policies π, for all instances $\nu \in \mathcal{E}$, we have:*

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

Proof. The essence of this proof, as it is in the original document is to focus on bounding the limit inferior of the expected number of samples from sub-optimal arms by consistent policies. We want to show in an instance $\nu \in \mathcal{E}$ that for all $i \in \mathcal{K}$, except

for i^* the optimal arm:

$$\liminf_{n\to\infty} \frac{E[T_i(n)]}{\ln(n)} \geq \frac{1}{D_{KL}(P_i \| P_{i^*})}\,.$$

As then the regret decomposition inequality yields our result trivially.

In the proof we will consider two instances ν and ν' in $\mathcal{E}_K$, and we will prove the result above for the first arm. In the first instance the first arm will have distribution P_1^ν with mean μ_1, and will be suboptimal, with the optimal arm being P_2^ν, with mean μ_2. In the second instance arm one will have $P_1^{\nu'}$ with mean $\lambda > \mu_2$, and all other arms will be identical to ν. We require some technical conditions to move forward with the proof, adapted from Lai and Robbins[19] for greater readability and less dense, updated notation.

We will require that the space $\mathscr{S}$ of distributions to which our arms belong is such that for any $\delta > 0$, and for any distribution P_λ in the space with mean μ_λ, there is a distribution P'_λ with mean μ'_λ such that $\mu_\lambda < \mu'_\lambda < \mu_\lambda + \delta$. This is the case for any family of distributions where the mean of any member belongs to a real interval, like a gaussian, exponential or poisson distribution. Since it is a technical quirk which does not really influence the theorem, it is given here instead. This condition is very similar to another one, we require, for all P_λ, P_θ with means $\mu_\lambda > \mu_\theta$, and for all $\epsilon > 0$, there is $\delta > 0$ such that $|D_{KL}(P_\theta \| P_\lambda) - D_{KL}(P_\theta \| P'_\lambda)| < \epsilon$ when $\mu_\lambda \leq \mu'_\lambda \leq \mu_\lambda + \delta$. This is simply continuity of the divergence over $\mathscr{S}$. We will require one final condition, that for arms i, and j with $\mu_i > \mu_j$ the Kullback-Leibler divergence from j to i is neither 0 nor infinite, i.e. that no arms are identical and that they do not violate absolute continuity conditions.

Our above requirements, allow us to choose $P_1^{\nu'}$ such that $\lambda > \mu_2$ and such that the following holds:

$$|D_{KL}(P_1^\nu \| P_1^{\nu'}) - D_{KL}(P_1^\nu \| P_2^\nu)| < \delta D_{KL}(P_1^\nu \| P_2^\nu)\,. \tag{2.1}$$

Since we are restricted to consistent policies, we have, for some $0 < \alpha < \delta$:

$$\mathbb{E}_{\nu'}[n - T_1(n)] = \sum_{i\neq 1} \mathbb{E}_{\nu'}[T_n(i)] = \mathcal{O}(n^\alpha)\,.$$

Now, picking up steam we give successive lower bounds:

$$\mathbb{E}_{\nu'}[n - T_1(n)] \geq \mathbb{E}_{\nu'}\left[n - \mathbb{I}\left\{T_1(n)\ln(n) \leq \frac{(1-\delta)\ln(n)}{D_{KL}(P_1^\nu \| P_1^{\nu'})}\right\}\right]$$

$$\begin{aligned}\mathcal{O}(n^\alpha) &\geq \mathbb{E}_{\nu'}\left[n - \mathcal{O}(\ln(n))\mathbb{I}\left\{T_1(n) \leq \frac{(1-\delta)\ln(n)}{D_{KL}(P_1^\nu \| P_1^{\nu'})}\right\}\right] \\ &\geq \mathbb{E}_{\nu'}\left[(n - \mathcal{O}(\ln(n)))\mathbb{I}\left\{T_1(n) \leq \frac{(1-\delta)\ln(n)}{D_{KL}(P_1^\nu \| P_1^{\nu'})}\right\}\right] \\ &\geq (n - \mathcal{O}(\ln(n)))P_{\nu'}\left(T_1(n) \leq \frac{(1-\delta)\ln(n)}{D_{KL}(P_1^\nu \| P_1^{\nu'})}\right)\end{aligned}$$

Now, define $L_m := \sum_{t=1}^m f(X_t; P_1^\nu)/f(X_t; P_1^{\nu'})$, the empirical divergence estimate where X_t are observed rewards from arm 1, with $m \leq T_1(n)$, of course. Then, relate both instances ν, ν' by defining the event:

$$C_n := \left\{T_1(n) \leq \frac{(1-\delta)\ln(n)}{D_{KL}(P_1^\nu \| P_1^{\nu'})}, L_{T_1(n)} \leq (1-\alpha)\ln(n)\right\}.$$

One can see that $P_{\nu'}(C_n) = \mathcal{O}(n^{\alpha-1})$. We have shown the first part of C_n to be $\mathcal{O}(n^\alpha)$, we can see that the second part, while dependent, reduces the probability by at least a factor of $\frac{1}{n}$. Further scrutiny highlights that C_n is a disjoint union of events of the form:

$$\mathcal{C} := \bigcap_{i=1}^k (\{T_n(i) = n_i\}) \cap \{L_{n_1} \leq (1-\alpha)\ln(n)\}.$$

Under constraints that:

$$\sum_{i=1}^k n_i = n \text{ and } n_1 \leq \frac{(1-\delta)\ln(n)}{D_{KL}(P_1^\nu \| P_1^{\nu'})}.$$

We want to now find a bound relating $P_{\nu'}(C_n)$ and $P_\nu(C_n)$. To do so, using the above formulation of C_n, notice that:

$$\begin{aligned}P_{\nu'}(C_n) &= \int_{T_1(n)=n_1,\ldots,T_n(k)=n_k, L_{n_1} \leq (1-\alpha)\ln(n)} \prod_{t=1}^{n_1} \frac{f(X_t; P_1^\nu)}{f(X_t; P_1^{\nu'})} \mathrm{d}\mathbb{P}_\nu \\ &= \int_{\mathcal{C}} \exp\left(\ln\left(\prod_{t=1}^{n_1} \frac{f(X_t; P_1^\nu)}{f(X_t; P_1^{\nu'})}\right)\right) \mathrm{d}\mathbb{P}_\nu \\ &= \int_{\mathcal{C}} \exp\left(-L_{n_1}\right) \mathrm{d}\mathbb{P}_\nu \\ &\geq \exp(-(1-\alpha)\ln(n)) \int_{\mathcal{C}} \mathrm{d}\mathbb{P}_\nu. \end{aligned} \tag{2.2}$$

Then, asymptotically, $P_{\nu'}(C_n) \geq n^{\alpha-1} P_\nu(C_n)$, as we desired. Now we present some asymptotical probability results related to C_n, which will bring us to the conclusion of our proof up to a small substitution.

$$\lim_{n\to\infty} P_\nu \left\{ \exists i \geq \frac{(1-\delta)\ln(n)}{D_{KL}(P_1^\nu \| P_1^{\nu'})} : L_i > (1-\alpha)\ln(n) \right\} = 0\,.$$

Then, combining with equation 2.2,

$$\lim_{n\to\infty} P_\nu \left\{ T_1(n) \leq \frac{(1-\delta)\ln(n)}{D_{KL}(P_1^\nu \| P_1^{\nu'})} \right\} = 0\,.$$

Finally, using our choice of $P_1^{\nu'}$ from equation 2.1, we have:

$$\lim_{n\to\infty} P_\nu \left\{ T_1(n) \leq \frac{(1-\delta)\ln(n)}{(1+\delta) D_{KL}(P_1^\nu \| P_2^\nu)} \right\} = 0\,.$$

Taking the limiting case for δ, as it can be arbitrarily small, and rearranging terms gives us:

$$\liminf_{n\to\infty} \frac{\mathbb{E}_\nu[T_1(n)]}{\ln(n)} \geq \frac{1}{D_{KL}(P_1^\nu \| P_2^\nu)}\,.$$

This has proven the theorem for an arbitrary sub-optimal arm, specifically 1, thus it holds for all sub-optimal arms in all environments, and combining with the regret decomposition identity, as stated at the beginning of the proof, gives the regret bound. □

This theorem only holds for single parameter distributions, but was extended to the multi-parameter case by Burnetas and Katehakis[8]. 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 ([23]). *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 distri-*

butions, 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. The proof was also our first intense foray into the details for which we developed tools in chapter 1. It is somewhat different from the rest of the proofs in this thesis, as it maintains the older methods of Lai and Robbins[19], while updating the contents and expanding on them to facilitate understanding by the reader. Several other proofs exist, such as the one by Bubeck et al.[7], or the one by Lattimore and Szepesvári[23], but both restrict themselves to a sub-class of bandit problems, and prove respectively corollaries 2.2.3 and 2.2.2 only, instead of the general case. The general case proof as given is more complex, and most likely the most complex proof in this thesis, but it is rewarding in the elegance of its result. The next section will give a finite time quantity to use as a 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 $\mathcal{E}_K$ 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. To prove this bound we must first introduce a new framework for describing the stochastic bandit setting, and prove a lemma on the KL-divergence.

This new canonical bandit model is a new angle from which to view the theory of multi-armed bandits. It is not the most obvious, but it can be very useful, and will be used in our proof of the minimax bound, thus we introduce it here. First we must define a new measurable space, in terms of a notion of *history*. It is natural, therefore, to begin with the following definition.

Definition 2.2.2. [History] The collection of all past actions and rewards is called the history of the problem. There are several ways to consider the history, we give here three notations to clearly differentiate them. We denote the random variable $H_n := (A_1, X_1, \ldots, A_n, X_n)$, and a realisation $h_n := (a_1, x_1, \ldots, a_n, x_n)$. We are also

interested in the sample space to which all H_n belong, the *history space*, which we denote $\mathcal{H}_n$. See that this sample space is clearly composed of n pairs each consisting of an integer less than K, the arm corresponding to the action, and a real valued reward. Thus, formally $H_n := (\llbracket K \rrbracket \times \mathbb{R})^n$. Do note, that this definition implies we henceforth hold K fixed.

To work in the history space $\mathcal{H}_n$, we would need to extend it to a measurable space, so that we can define measures rigorously in $\mathcal{H}_n$. To do so, we will use the Lebesgue σ-algebra, based on the Lebesgue measure mentioned in section 1.1.1. The Lebesgue σ-algebra on S, $\mathcal{L}(S)$ is the σ-algebra of all sets in S measurable with respect to the Lebesgue measure. In our case, we take $\mathcal{L}_n := \mathcal{L}(\mathcal{H}_n)$, which is equivalent to restricting the Lebesgue sigma-algebra of R^{2n} to $\mathcal{H}_n$.

See that now, a single realisation from $\mathcal{H}_n$ gives us the entirety of the contents of an n round run against a K-armed bandit. It is easily seen in fact that $\mathcal{H}_n$ consists precisely of any possible course of proverbial events in an n-round, K-armed, bandit. From this realisation we can see that the interaction of any instance $\nu \in \mathcal{E}_K$ and any policy π is in $\mathcal{H}_n$. Since we now have a valid measurable space, the next logical step is to look at measures on this space. Consider the random variables A_t, X_t for $t \leq n$, on the canonical bandit $(\mathcal{H}_n, \mathcal{L}_n)$, see that for all $h_n \in \mathcal{H}_n$, 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 still, they do not depend on an instance or policy.

In this new canonical model, we now set out to prove 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.

Lemma 2.2.4 ([24]). *In a K-armed bandit instance ν, with arms with distribution P_i for $i \in \llbracket K \rrbracket$, 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 $P_\pi(a_t | a_s, x_s : s \leq t)$ 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.

Proof. Expanding, by conditioning, we have:

$$\begin{aligned}
\mathrm{d}\mathbb{P}_\nu(h_n) &= \mathrm{d}\mathbb{P}_\nu(h_n | h_{n-1}) \mathrm{d}\mathbb{P}_\nu(h_{n-1}) \\
&= \prod_{t=1}^{n} \mathrm{d}\mathbb{P}_\nu(h_t | h_s : s < t) \\
&= \prod_{t=1}^{n} \mathrm{d}A_t(h_t | h_s : s < t) \mathrm{d}X_t(h_t | h_s : s < t)
\end{aligned}$$

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

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:

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

□

Lemma 2.2.5 ([24]). *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:

$$D_{KL}(\mathbb{P}_\nu \| \mathbb{P}_{\nu'}) = \sum_{t=1}^{n} \mathbb{E}_\nu\left[D_{KL}(P_{A_t} \| P_{A_t})\right]$$

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

□

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 [24]). *The minimax pseudo-regret of the class $\mathcal{E}_K$, given $n > K - 1$ is:*

$$\bar{R}_n^*(\mathcal{E}_K) \geq \frac{1}{27} \sqrt{n(K-1)} \,.$$

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 $\mathcal{G}_K$, rather than $\mathcal{E}_K$. This slight lack of rigour is justified by the assumed sub-gaussianity of rewards in $\mathcal{E}_K$. Intuitively, if the arms have lighter tails than a gaussian distribution, then distinguishing arms should be easier, meaning that the minimax regret is presumably higher over the class $\mathcal{G}_K$ than of bandits with stricktly lighter reward tails.

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 performing π. We begin with the first environment. Let $0 \leq r \leq \frac{1}{2}$, the first environment in $\mathcal{G}_K$ has means r for arm 1 and 0 otherwise. In this environment we denote $\mathbb{P}_1$ the distribution on the canonical bandit $(\mathcal{H}_n, \mathcal{L}_n)$.

For the second environment we take the optimal arm i to be arm least taken by π in 1: $i = \text{argmin}_{j \neq 1} \mathbb{E}_1[T_j(n)]$. 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 rewrite both regrets $\bar{R}_n^1(\pi)$ and $\bar{R}_n^2(\pi)$, then apply the Pinsker-type inequality (theorem 1.2.1) and finally lemma 2.2.5.

$$\bar{R}_n^1 + \bar{R}_n^2 > \frac{nr}{2} \left(\mathbb{P}_1(T_1(n) \leq \frac{n}{2}) + \mathbb{P}_2(T_1(n) > \frac{n}{2}) \right)$$

$$> \frac{nr}{4} \exp\left(-D_{KL}(\mathbb{P}_1 \| \mathbb{P}_2)\right) . \tag{2.3}$$

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 2.3 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. □

This theorem gives us guarantees on the worst-case regret of any algorithm over any possible stochastic bandit environment. In many instances the regret will be less than this bound, but unavoidably, some instances 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 this algorithm might perform terribly on nicer instances, where logarithmic asymptotic growth may be reached well 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. Regret properties will be given, some simple ones proven, others taken from the literature, and will be compared 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.[32]. 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 that we explore until the mean of each arm is known with sufficient confidence relative to some parameter. While this second strategy may seem more ef-

fective, and indeed it might be, both strategies will suffer from the same fundament 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 2.1.1. First, we shall present a general formula for the regret of the ETC strategy. After discussion 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[22])**.** *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 2.1.4, 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 necessary 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, if they were there would be no need for a bandit. 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:*

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

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 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? Corrollary 3.1.2 suggests we are still quite far from optimality, but this could be due to our analysis. Unfortunately, this is not the case. While more complicated and tighter bounds can be derived, Garivier et al.[13] showed that even policies which are optimal within the class of ETC strategies only achieve asymptotic growth of $\ln(n)/\Delta$, which is already half that of fixed design strategies, but the algorithms presented in the next section can achieve half that. If the sub-optimality gap is unknown, however, explore-then-commit strategies suffer terribly as there is no good way to choose m, and in fact in the general case there is no good way to determine a stopping time τ. This causes them to have an un-improvable regret of $\mathcal{O}(n^{2/3})$[30][1]. In comparison to the minimax regret $\Omega(\sqrt{nK})$, this is a catastrophic gap in the domain where $K \leq \sqrt[3]{n}$.

This 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. Use of heuristics such as the *doubling trick* cannot circumvent this issue, the only solution to achieve optimal behaviour is to design a fully sequential strategy which constantly evaluates wether 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

[1] Perchet et al.[32] also credit Somerville for this result, but the author could not verify this prior claim.

algorithm is to use an index for each arm and play the arm with highest index each turn until the end of the game.

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 is in fact what will allow us to always asymptotically choose the right arm, unlike ETC.

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[19] and so named by Auer et al.[4]. The algorithm we present and refer to as *UCB* algorithm is given by Lattimore and Szepesvári[27]. 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 2.1.4), 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 $\delta^{-1} = f(t) := 1 + t\log^2(t)$ 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[27])**.** *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. *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 2.1.3, 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 2.1.1), 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 chose 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.

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

The above follows from redefining 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 $\hat{\mu}_{i,s} = \frac{1}{s}\sum_{j=1}^{s} Z_{i,j}$. Now that we have weeded $T_i(n)$ out of our mean, so we can move to again applying theorem 2.1.3:

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

Rearranging S_2 into a form to which we can apply lemma 3.2.2 yields:

$$
S_2 = \sum_{t=1}^{n} \mathbb{I}\left\{ \hat{\mu}_i(t-1) + \sqrt{\frac{2\ln f(t)}{T_i(t-1)}} \geq \mu_o - \epsilon, A_t = i \right\}
$$

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

Combining S_1 and S_2, and completing the regret decomposition identity, leads to the desired result. The infimum insures 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:

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

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

The bonds of theorem 3.2.1 can be simplified trivially by a clever choice of ϵ, which is given in corollary 3.2.3.

Corollary 3.2.3 (UCB Regret Bounds (Simplified)[27]). *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. 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[23])**.** *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*. To the best knowledge of the author, there are no known policies which are both minimax optimal and asymptotically optimal, like there are no algorithms that are perfectly asymptotically optimal for all classes of problems. 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 order, but achieves $\mathcal{O}(\sqrt{nK\ln(K)})$ regret in the worst case, which is notably tighter.

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

$(\mathcal{H}_n, \mathcal{L}_n)$ The canonical bandit (a measurable space)

$\bar{R}_n$ 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 $\cdot$

$\mathcal{E}$ 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

$\mathcal{H}_n, H_n, h_n$ Histories in the Canonical model, see definition 2.2.2

$\mathcal{K}$ The arm set

$\mathcal{O}(\cdot), \Omega(\cdot)$ "Big-O" order notation

μ^* The mean of the optimal arm

ν, ν' Stochastic bandit instances

$\perp$ Independence

π A policy

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

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

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 $\mathbb{R} \cup \{-\infty, \infty\}$
$f(\cdot)$	A density, defined as a Radon-Nikodym derivative

Miscelaneous

:	"such that"

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] Gittins, J. C.
1979. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177.

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

[18] Kveton, B., Z. Wen, A. Ashkan, H. Eydgahi, and M. Valko
2014. Polymatroid bandits. *CoRR, abs/1405.7752*.

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

[20] Lattimore, T. and C. Szepesvári
2016a. Bandits: A new beginning. `http://www.banditalgs.com/2016/09/04/bandits-a-new-beginning/`.

[21] Lattimore, T. and C. Szepesvári
2016b. Finite-armed stochastic bandits: Warming up. `http://banditalgs.com/2016/09/04/stochastic-bandits-warm-up/`.

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

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

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

[25] Lattimore, T. and C. Szepesvári
2016f. Optimality concepts and information theory. http://banditalgs.com/2016/09/22/optimality-concepts-and-information-theory/.

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

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

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

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

[30] 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.

[31] Minsky, M.
1961. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30.

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

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

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

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

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

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

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

[39] 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.

[40] 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.

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

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

[43] 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.