Mathematics & Statistics | Lancaster University

Multi-Armed Bandits

Sequential Decision Agents in Reinforcement Learning

by

Lorenzo Croissant

Supervisor:

Dr. Azadeh Khaleghi

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

June 2018

Contents

Consider the problem of a clinical trial where several new drugs are tested to replace an existing drug. The most simple experimental framework would be to assign a fixed number of patients to the new drugs. As new patients join the experiment they are allocated to one of the drugs until each has the required number of patients. If one of the drugs can be ruled to be ineffective with relatively high confidence early in the test, shouldn't we stop assigning patients to it in favour of the others? The key medical ethic "do no harm" would suggest that we should seek to find the most effective flexible procedure to ensure we do not needlessly put patients at risk in clinical trials.

Chapter 1

Measure and Information Theory

1.1 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 up to 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[8].

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 in particular 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 $E, F \in \mathcal{E}$: $E \cup F, E - F \in \mathcal{E}$. 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 is not restricted to them at all. A particularly important example 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-field, which we denote $\mathcal{B}$. $\mathcal{B}$ 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 $e \in \mathcal{E}$, $M(e) \in S$. In particular 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.

In particular, perhaps the most important measure 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 (Measurablitiy)**.** Let $(\mathcal{X}, \mathcal{S})$ be a measurable space, i.e. $\mathcal{S}$ is σ-field 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-field, 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-field $\mathcal{B}^*$ 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:

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

In particular if f is real valued we can take $(\mathcal{Y}, \mathcal{T}) := (\mathbb{R}^*, \mathcal{B}^*)$ and obtain the following condition:

$$\forall B \in \mathcal{B}^* : f^{-1}(B) \in \mathcal{S} .$$

Notice that the probability measure is a map from the set of all possible events in $\mathcal{X}$, the σ-field $\mathcal{S}$, 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. This is usually illustrated by thinking of a random variable $X(\omega)$ not containing any inherent randomness, but instead deterministically mapping (as a probability measure) from the random outcomes $\omega \in \Omega := \mathcal{X}$ to the interval $[0, 1]$. 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 The Radon-Nikodym theorem

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, s is said to hold almost everywhere if it holds for a set $A \in \mathcal{S}$ such that $\mu(A^c) = 0$. 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. *If f is integrable with respect to a measure μ, it is finite almost everywhere with respect to μ.*
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)**.** *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. *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)**.** *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 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[9], 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[6]. This is because we do not truly need 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 Feinstein[5], or Shannon[9] 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 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 experiment, 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 the event $\{X(\omega) = x\}$ for differentiating H_1 and H_2. We can then define the mean information using the generalised probability densities f_1, f_2 for $\mu_1(E) \neq 0$:

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

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 KL-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 concept will be bounding a random variable called the *pseudo-regret*, and for one such bound we will require a class of results known as the *Pinsker-type inequalitiesgt*.

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. In particular, we want to look for 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). *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, with Radon-Nikodym derivatives with respect to μ f_1 and f_2 respectively. 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 take the logarithm into the integral:

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

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 reformulate probability theory in terms of topological spaces, σ-fields 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 discrete 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 a new for of continuity and uniqueness. These ideas will be used throughout this thesis, and in particular 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 *stochastic bandit* problem, we further restrict study to the case of a stationary environment, with K arms, which form a set $\mathcal{K}$, each with an underlying distributions from which rewards are drawn independently at random, until a fixed time horizon N. There are many further assumptions to be made in this precise problem, for example that rewards are observed without bias or delay. Many different sets of assumptions have been studied in stochastic bandits, which relax one or several of these assumptions, or adding further assumptions to extract higher performance in specific settings. Notwithstanding, we restrict ourselves here to the simplest case, and encourage the reader to explore further resources should they wish to challenge the assumptions of this framework. In this chapter, we will establish some mathematical foundations which we will use to evaluate the performance of two algorithms which we will study. While the first, explore-then-commit is an intuitive strategy, it has many flaws which will highlight the difficulty of bandit problems, and will motivate the more advanced upper confidence bound algorithm.

2.1 Mathematical notes

THE main element of mathematical interest in this thesis is the concept of regret. Like in 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 very important later. 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[4]:

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

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.

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:

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

□

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:

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

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

□

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 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 as $N < \infty$. 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)**.** *Let X be a sub-gaussian random variable.*

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

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

  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 $\frac{t^2}{2} > 0$, 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)**.** *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. In this next section we will design and review our first algorithm for competing in the stochastic bandit environment.

2.2 Explore-Then-Commit

FOLLOWING these mathematical notes, we have the tools required to begin designing and evaluating algorithms. Recalling the stochastic bandit problem, we begin by designing a simple and intuitive strategy, and then analyse its regret performance. What is the simplest strategy that comes to mind? Intuitively we need to do two things in sequence: find the best arm, repeatedly exploit it. Why not simply explore for a given number of rounds $M < N$, then take the arm with highest mean and play that arm the remaining $N - M$ rounds? This strategy is valid, and is called the explore-then-commit strategy.

2.2.1 Algorithm

Before introducing the explore-then-commit (ETC) algorithm we need to clarify a few details. First, we want our exploration to be uniform over the arms, thus we will choose a deterministic exploration and take $M = mK$, for some $m \in \mathbb{N}$. Second, we must consider the case of a tie for best arm at round M, where we will break the tie in a deterministic way, say by choosing the arm of least index. Finally, we will assume that the noise in the rewards of our stochastic bandit is σ^2-subgaussian, with σ^2 known. Throughout this thesis we will study cases of 1-subgaussian random noise, for simplicity, but the results will hold for all valid σ^2. The formal algorithm is presented in Algorithm 1

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

Algorithm 1: Pseudo-code for the ETC algorithm

ETC is a simplistic strategy, but as an intuitive solution to the stochastic bandit problem it will help us introduce regret bounds using the tools from section 2.1, and better understand why more flexible approaches can offer better solutions to this bandit problem.

2.2.2 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 2.2.1 (General ETC regret)**.** *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 form 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 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. An instance where we can consider the gap known, and also the simplest stochastic bandit is the two-armed bandit. Here, one can take arm 1 to be optimal, and arm 2 to have gap Δ, as there is only one gap.

Corollary 2.2.2 (ETC regret for two-armed bandits)**.** *In the case of a two-armed stochastic bandit with* 1*-subgaussian noise, the ETC algorithm satisfies:*

$$\bar{R}_n \leq \frac{\Delta}{2}\left(1 + \ln \frac{n\Delta^2}{2}\right) \; .$$

Proof. We apply theorem 2.2.1 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:

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

□

2.3 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. Note that in ETC, while exploring we sampled all arms equally. This is not necessary as if the confidence intervals for two arms don't overlap, there is good reason to simply abandon the lower one. Thinking about the bounds of intervals instead of their centre is the key to understanding the Upper Confidence Bound (UCB) algorithm. This algorithm values the highest upper bound on the arm means rather than the highest estimate, in this sense it is an optimistic algorithm.

2.3.1 Algorithm

An interesting fact the reader might note right away is that in the UCB algorithm as it has been simply presented above it is entirely possible for the preferred arm to change over time, which takes a two stage approach to exploration and exploitation off the table. We will explore and exploit simultaneously, but to insure our algorithm performs well we will add to it a component which forces us to explore arms which have not been played often. This may seem like a burden, but is in fact what will allow us to always asymptotically choose the right arm, unlike ETC. Before writing out the algorithm our first task is to find this tempered confidence interval and explain it.

Recall Hoeffding's bound, which implies that for ϵ we have $P(\hat{\mu} \geq \epsilon) \leq \exp\left(\frac{n\epsilon^2}{2}\right)$. Letting $\delta := \exp\left(\frac{n\epsilon^2}{2}\right)$, we have $P\left(\hat{\mu} \geq \sqrt{\frac{2}{n}\ln\left(\frac{1}{\delta}\right)}\right) \leq \delta$. This is a plausible interval which we can adjust using the parameter δ, which requires us to assume that the $X_i - \mu$ are sub-gaussian. We can now deduce the smallest plausible (w.r.t. δ) upper bound for $\hat{\mu}$ to be

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

We can continue this thread and choose a convenient δ, 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 in case we mistakenly disregard the best arm. We will not discuss this until the next section, however, it can be seen that we require $\delta^{-1} = f(t) := 1 + t\log^2(t)$. We are now ready to introduce the UCB algorithm, whose pseudo-code is included in algorithm 2.

```
Input: $0 < \delta < 1$
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

It is important to note that while algorithm 2 is referred to as *the* UCB algorithm in this thesis, UCB algorithms are a family, of which this algorithm is simply an example. More general formulations take several arguments[4], and expand the range of confidence bounds and exploration bonuses used. We will analyse the regret only of this formulation of UCB, but more general results exist.

2.3.2 Regret Analysis

Using the tools developed in section 2.1 we will now prove several results about the regret of the UCB algorithm. We focus mostly in this section on instance dependent worst-case regrets, beginning with theorem 2.3.1, which gives a bound on the regret and an asymptotic bound on the factor of logarithmic growth of the regret. These bounds together give us a good idea of the behaviour of the regret for the two most relevant domains of n.

Theorem 2.3.1 (UCB Regret Bounds)**.** *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. $\limsup_{n \to \infty} \frac{\bar{R}_n}{\ln(n)} \leq 2 \sum_{i:\Delta_i > 0} \frac{1}{\Delta_i}$

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*. The astute reader might

Lemma 2.3.2. *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 seperately 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 2.3.2 yields:

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

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

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:

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

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

□

The bonds of theorem 2.3.1 can be simplified trivially by a clever choice of ϵ, which is given in corollary **??**.

Corollary 2.3.3 (UCB Regret Bounds (Simplified))**.** **??** *Choosing* $\epsilon = \frac{\Delta_i}{2}$ *in theorem 2.3.1 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:*

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

The second bound is particularly interesting, as it shows that the logarithmic growth of the regret of UCB is controlled by the sum inverse of the sub-optimality gaps, without the notion of limit superior outlined in theorem 2.3.1. To complete this overview of the regret of UCB, we will derive its order, in theorem 2.3.4.

Theorem 2.3.4 (Order of UCB Regret)**.** *The pseudo-regret of an 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 Δ, gives $\mathbb{E}[T_i(n)] \leq C\ln(n)\Delta_i^{-1}$ from which we obtain a distribution free bound:

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

Letting $\Delta = \sqrt{K\ln(n)n^{-1}}$ gives the result.

□

2.3.3 Minimax Regret

The *minimax* regret is a fundamental quantity of the difficulty of the stochastic bandit problem. 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 regrets, but if we are in the worst possible instance, the minimax is a meaningful lower bound for the best performance of our algorithm.

Theorem 2.3.5. *The worst case pseudo-regret of any bandit in 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. 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

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 $\bar{R}^1_n(\pi)$ and $\bar{R}^2_n(\pi)$, then apply the Pinsker-type inequality (theorem 1.2.1) and finally the divergence decomposition lemma.

$$\begin{aligned}\bar{R}^1_n + \bar{R}^2_n &> \frac{nr}{2}\left(\mathbb{P}_1(T_1(n) \leq \frac{n}{2}) + \mathbb{P}_1(T_1(n) > \frac{n}{2})\right) \\ &> \frac{nr}{4}\exp\left(-D_{KL}(\mathbb{P}_1 \| \mathbb{P}_2)\right) . \qquad (2.1)\end{aligned}$$

Now, we use the divergence decomposition lemma 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 2.1 yields:

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

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

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

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:

$$\begin{aligned}\bar{R}_n^* &\geq \frac{29}{768}\sqrt{n(K-1)} \\ &\geq \frac{1}{27}\sqrt{n(K-1)}\,.\end{aligned}$$

In the last line, $\frac{1}{27}$ is the largest simple fraction less than the fraction on the previous line. This change only tidies up the formula. □

2.3.4 Lower Bounds

Bibliography

[1]
.

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

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

[4] 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.

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

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

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

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

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

[10] 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.