Pure Exploration in Multi-Armed Bandits

by

Connor James Stephens

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistical Machine Learning

Department of Computing Science

University of Alberta

Abstract

Many practical problems in fields ranging from online advertising to genomics can be framed as the task of selecting the best option from among several choices, based on a limited number of noisy evaluations of the quality of each choice. Pure exploration in multi-armed bandits is an online-learning setting which aims to capture the challenge of designing adaptive data collection and terminal selection procedures that optimize the quality of their final recommendation. In this thesis we provide a comprehensive overview of the sample complexities, as well as algorithms that achieve them, for several core pure exploration settings. We cover problems in both the fixed confidence and fixed budget frameworks and provide a number of novel contributions, including a new analysis of a well-known algorithm, Sequential Halving. This result reveals new performance guarantees and contributes to a greater understanding of the sample complexity of making close to optimal selections in the fixed budget setting, a highly practical problem setup which has remained largely overlooked in the literature until recently.

Preface

Section 3.4 and Section 4.2.1 contain original contributions of the author, developed through discussions with my supervisor, Csaba Szepesvári. Section 4.3 of this thesis contains results obtained in collaboration with Yao Zhao and Kwang-Sung Jun at the University of Arizona as well as Csaba Szepesvári. The proofs for the upper bounds for Sequential Halving in Section 4.3 were primarily developed by our co-authors, while the development of the lower bound for returning sets of good arms with uniform sampling presented in this work (Theorem 4.17) was a joint effort between Kwang-Sung Jun and the author. At the time of writing extensions of this work are being prepared for submission to the fortieth International Conference on Machine Learning (ICML 2023). The figures in this thesis are best viewed in colour.

“No regrets.”

– Bandit algorithms, probably.

Acknowledgements

To start off I would like to thank my supervisor, Csaba Szepesvári, for his support and for the thoughtful conversations that went into this work. I am grateful to have had this opportunity to work with Csaba, and I deeply appreciate the trust and freedom that I was granted to find my own way. Csaba, I forgive you for pointing out that I do not know the difference between 'which' and 'that'. This is likely only the tip of the iceberg.

I would also like to thank some of the people who helped me to reach this point. It is not an exaggeration to say that Andrzej Czarnecki changed the course of my life as a researcher. Andzrej wrote me his personal phone number on a torn-out page of his passport, and his support has been the very definition of unwavering ever since. Frank Marsiglio's door was always open, his office always smelled like coffee, and he encouraged me to pursue my research interests outside of physics. Martin Jagersand made the big leap possible. Martin lent me a great deal of his own time to help me to get to where I wanted to be. I am grateful for all of these people.

I am also grateful for the friends that I have made over the past few years. I'm glad that I was able to be a part of that first warehouse of new graduate students, comparing hottest hometowns and performing awful karaoke. It has been a pleasure to be a part of the cheerfully chaotic RLAI lab for the past two years; there is no place that I would rather not get work done. I would like to thank Alex, Amir, Andy, Laura, Gabor, Chunlok, Prabhat, Vlad, Mikhail, Matt, Brad, Aidan, and others that I have likely missed here. Finally, I would like to thank my family, for getting me to the finish line.

I would also like to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 A Brief History of Pure Exploration in Bandits . . . 3
- 1.3 Current Research Directions . . . 5
- 1.4 Contributions . . . 6

2 Background **8**
- 2.1 The Objective(s) . . . 9
- 2.2 Additional Notation . . . 11
- 2.3 Cumulative Regret Minimization versus Pure Exploration . . . 12
- 2.4 Finite-Time Concentration Inequalities . . . 13
- 2.5 Information-Theoretic Lower Bounds . . . 14

3 Pure Exploration with Fixed Confidence **21**
- 3.1 PAC Identification of Good Arms . . . 21
- 3.2 Best Arm Identification . . . 26
- 3.3 Making Sense of Many Good Arms: Alternative PAC Frameworks . . . 36
- 3.4 Aside: Instance-Optimal Rates for Stochastic Convex Optimization . . . 42

4 Pure Exploration with a Fixed Budget **47**
- 4.1 Best Arm Identification . . . 48
- 4.2 Simple Regret . . . 55
- 4.3 Faster Rates for Identifying Good Arms . . . 68
- 4.4 Anytime Algorithms via the Doubling Trick . . . 75

5 Conclusion **77**
- 5.1 Summary . . . 77
- 5.2 Future Directions . . . 79

References **81**

Appendix A Proofs for Chapter 2 **85**
- A.1 Proof of Lemma 2.6 . . . 85

Appendix B Proofs for Chapter 3 **91**
- B.1 Proof of Theorem 3.16 . . . 91

Appendix C Proofs for Chapter 4 **93**
- C.1 Proof of Theorem 4.2 . . . 93
- C.2 Proof of Theorem 4.10 . . . 96
- C.3 Proof of Proposition 4.16 . . . 98
- C.4 Proof of Theorem 4.17 . . . 99

C.5 Proof of Theorem 4.15 102

List of Figures

2.1 A sample voting selection for the New Yorker caption contest (`https://www.newyorker.com/cartoons/vote`) on the week of September 25th 2022. Users vote on the quality of the caption served to the user by the website. Given that each user will have their own preferences over captions we can treat an assessment of caption quality from a given user as a noisy sample of the true population appraisal of the caption. 10

3.1 An illustration of the behaviour of Uniform Exploration with Adaptive Stopping (Algorithm 2). In a) we see the sample means and confidence intervals after t 'chunks' of uniform sampling ($n \cdot t$ total samples), noting that they are of equal widths due to our vector-at-a-time sampling procedure. Here the intervals for arms 1 and 2 overlap and so we take another sample from each arm. In b), after $t + 1$ rounds of uniform sampling the sample means and confidence intervals have shifted, but a small overlap remains. In c), after $t + 2$ rounds of uniform sampling arm 1 has completely separated from the rest of the arms. At this point Algorithm 2 would stop and return arm 1. 28

3.2 When each arm i's confidence interval is no wider than Δ_i (taking $\Delta_1 \coloneqq \Delta_2$) then the unique best arm's confidence interval is disjoint from all other intervals. On an instance with uniform gaps such as in a) this requires uniformly many samples across all arms, whereas for a general instance as in b) we see a larger variation in confidence interval width (and thus numbers of samples required to produce said intervals). 30

3.3 An illustrative example of sets of (m, ε)-good arms for different values of m, ε on the same instance. Arms with blue stems are (m, ε)-good while arms with orange stems are not. A pair m, ε defines a threshold on arm means, illustrated here by a dashed grey line. 38

3.4 Examples of function f and adversarial chosen functions g chosen so that their ε-optimal regions have no overlap but the subgradients of the functions are are close as possible on the domain. The upper figure shows the relatively straightforward case for quadratic functions $f(x) = a(x - c)^2/2$ for some $a > 0, c \in \mathcal{D}$, and the lower figure shows an example for the generalized absolute value function class $f(x) = \frac{1}{k}|x - b|^{1/k}$ for $k \geq 1, b \in \mathcal{D}$. 45

4.1 Pure exploration performance of the $\mathcal{O}_\gamma$ oracle allocation family for $n = 1000$, averaged over 100,000 replications with $T = 3n \log n$. With such a small sample budget the probability of selecting a mean zero arm dominates the simple regret and vanishingly small values of γ are optimal. In contrast, due to the small gap between the top two arms a constant fraction of samples needs to be allocated to the competitor arm for optimal BAI error (approaching 1/3 from below in the limit as n increases). 64

4.2 The pair $\tilde{m}, \tilde{\varepsilon}$ that minimizes the bound in Corollary 4.15.1 roughly correspond to the pair that maximizes the quantity $\tilde{m} \cdot \tilde{\varepsilon}^2$ while satisfying $\Delta_m + \tilde{\varepsilon} < \varepsilon$. We show examples of such pairs on two different instances. In each of these figures the arms with blue stems are ε-good, and arms with orange stems are ε-bad. The improvement in the sample complexity guarantee from the näive choice of $\tilde{m} = 1, \tilde{\varepsilon} = \varepsilon$ on the instance in a) is $\tilde{m} \cdot \tilde{\varepsilon}^2 = 3\varepsilon^2$ and in b) is $\tilde{m} \cdot \tilde{\varepsilon}^2 = \frac{25}{16}\varepsilon^2$. . 72

Chapter 1

Introduction

1.1 Motivation

Imagine that your friend has just started a small online clothing store where they plan to sell elaborately embroidered sweaters. Setting up production for a new sweater design is a time and labour intensive process, involving designing the new embroidery pattern, setting up the sewing equipment to stitch the pattern, and obtaining the material required for the design. Due to the small-scale nature of their business and the limited amount of sweaters that they can produce each day they plan to maximize their return on initial production set-up time and cost by making a large batch of one design at a time over several weeks.

They have several sweater designs that they are considering for production and want to determine which design will be the most popular and result in the most sweaters sold. In order to find out they pay to have a mini-questionnaire displayed on websites: each questionnaire shows a user a *single* sweater design and asks them to rate their interest in buying the sweater on a scale from one to three. Your friend is charged a small fee by the hosting service each time a user interacts with a questionnaire and so in order to keep costs within their budget they pay the hosting service for 500 responses, after which they will commit to a design based on the results. To paraphrase the situation: your friend wants to use a fixed number of samples of consumer preference to choose a sweater that will optimize some statistic of the number of sweaters sold. For example they could look to maximize the expected number of sweaters sold, or the probability that they end up selecting the most popular sweater.

If your friend was in charge of the way that designs were shown to users they may choose a traditional A/B/n testing approach which would be to show each user a sweater selected

uniformly at random. A disadvantage of this approach is that if some of the designs are truly far more popular than others then this may not be the most efficient use of their sampling budget; a more powerful approach could be for them to take a sequential decision making approach towards collecting information, for instance ruling out designs that users have already indicated are clearly unpopular and by doing so reserving more samples to make a better selection from among the remaining designs.

This scenario has several characteristics which are essential to the problems considered in this thesis:

- A selection among a finite collection of options has to be made based on stochastic feedback about the quality of each individual option.
- The quality of each option is assumed to be fixed during the sampling process (e.g. we assume that fashion trends will not change during the time-span of our user survey), and each sample of this quality is assumed to be independent and identically distributed (i.i.d.).
- Samples can be collected in a fully adaptive fashion; the decision of which sweater to show the next user can be decided based on all previous observations, and in relevant scenarios the decision to stop sampling can also be made in a data-dependent manner.
- There is no 'cost' for taking a sample, other than sampling budget constraints or considerations of sample efficiency.

A multi-armed *bandit instance* is a finite collection of distributions, or 'arms'[1], and the task of sampling from these objects in an online fashion with the goal of making a selection among these arms is known as *pure-exploration in multi-armed bandits*. The phrase 'pure-exploration' refers to the idea that in these problems there is no cost to collecting samples apart from possible budget considerations, and the sole objective is to collect data in a manner which enables an optimal selection under uncertainty. Going back to the sweater example, we assume that asking a user to rate a given sweater design will not impact their likelihood of buying whichever sweater that we eventually decide to produce. In contrast, a clinical drug trial is a scenario in which assigning treatments to patients must necessarily trade off gaining information on the

[1]The name 'multi-armed bandit' is meant to evoke a slot machine with several arms to choose from.

quality of a given treatment with the administration of a possibly inferior treatment which has an explicit cost in terms of the quality of care provided to the patients in the trial. The pure exploration framework is a natural fit for settings with access to a simulator-like data collection process, for instance tuning supervised machine learning models on re-sampled or synthetic data, or simulated roll-outs of games, e.g. in the Monte Carlo tree search algorithm (Kocsis and Szepesvári, 2006). The goal of designing algorithms for pure-exploration in multi-armed stochastic bandits is to optimize the sample collection and recommendation process in order to make the best decision possible given a budget constraint, or alternatively to ensure that the selection has some minimum level of quality while using as few samples as possible.

1.2 A Brief History of Pure Exploration in Bandits

The pure exploration setting is the culmination of a series of increasingly more general and powerful statistical problem settings stemming from the classical analysis of variance (ANOVA) framework that underlies A/B tests. The classical statistical version of the pure exploration problem is concerned with a scenario in which a batch of samples has been collected uniformly from each of the arms of the bandit instance and some sort of decision needs to be made on the basis of these samples. Paulson (1952) is one of the earliest works of this form, and considers the problem of designing an optimal test to identify whether one of n normal distributions with common known variance has a larger mean than the rest, and if so, which one, essentially framing the problem of identifying a good option by way of an ANOVA test.

A major progression in the field appeared by Bechhofer (1954), in which it was argued that ANOVA was an unsuitable framework for the goals of experimenters who were interested in making an optimal decision among several options. Hypothesis tests are concerned with the detection of differences between populations under strict error control requirements, but most practitioners are only concerned with making a reasonable selection based on the data. Still considering a pre-drawn batch of samples, Bechhofer shifted from hypothesis testing towards producing a ranking of arms. His 1954 paper considered the smallest fixed per-arm sample size required to identify the arm with the largest mean with high probability under the assumption that the gaps between arm are larger than some threshold specified by the experimenter, a formalism which became known as the 'indifference-zone' approach.

While previous works focused on experiments with fixed sample sizes, initial forays into sequential sampling procedures arrived in the context of ranking and selection problems where the bandit arms have a common but unknown variance. In this setting no fixed-sample size procedure can guarantee the probability of making a correct selection, as the mistake probability necessarily depends on the variance which is not known at the time of sample size selection. Bechhofer, Dunnett, et al. (1954) proposed a two stage procedure to address this problem. It was not until the concurrent works of Paulson (1964) and the monograph Bechhofer, Kiefer, et al. (1968) that the use of sequential sampling procedures for the purpose of increasing the statistical power of procedures was explored. Bechhofer, Kiefer, et al. made use of a 'vector-at-a-time' sequential sampling paradigm, essentially combining uniform sampling of arms with a data-dependent stopping rule in a framework known as sequential analysis. The first appearance of a truly adaptive data collection and selection procedure was in Paulson's work in which he introduced a stage-wise elimination procedure to identify the arm with the largest mean under the indifference-zone assumption. While the form of the elimination procedure closely mirrors modern developments, notions of optimal arm elimination thresholds and sample complexity analyses for the proposed algorithms were left to later work.

By the late 1970s the asymptotic sample complexity of selecting the best arm under an indifference-zone was being considered. Jennison et al. (1982) was the first work to show that algorithms based on sequential analysis (i.e. algorithms which uniformly sample arms until an adaptive stopping condition) necessarily have a larger worst-case asymptotic sample complexity compared to fully adaptive sampling rules, and designed adaptive algorithms which have the correct worst-case asymptotic sample complexity as the prescribed error probability of the procedure approaches zero (again, in the indifference zone setting).

Modern research in pure exploration problems in bandits arguably began with Even-Dar et al. (2002), who applied the probably approximately correct (PAC) formalism to bandits. Their results were a large step toward modern research directions, providing a finite-time bound for the expected number of samples required to return a close-to-optimal arm with probability greater than some confidence threshold, and lower bounds that hold for non-asymptotic values of the confidence parameter.

1.3 Current Research Directions

In the previous section we saw how the problem settings being addressed in research have slowly moved towards increasingly practical and general questions. For example, in the previous section we briefly touched on the selection of a best arm with an indifference-zone, which is to say a problem setting in which we attempt to identify the arm with the largest mean, under the assumption that the best arm has a mean at least $\varepsilon > 0$ larger than the next arm. The motivation behind this problem setting is twofold: (a) the existence of a minimum gap makes the design and analysis of algorithms relatively straightforward and (b) if the top two arms had a smaller gap than ε then it was assumed that the error would be made between the top two arms and the practitioner would be satisfied with algorithms lacking error control in this case. by Jennison et al. (1982) the authors noted that while previous works had made progress on finding the largest mean with this assumption, with a small modification of the procedure it was also possible to reason about the probability of simply returning an arm with a mean no worse than the specified minimum gap without any assumptions on the gap structures of the arms. This second result provides a stronger result which better matches with the intentions of the practitioners running these algorithms. In a sense this broadly describes the direction of pure exploration research today: it is not enough to provide optimal algorithms for well specified problem settings, the problem settings we study should be practical, and result in useful algorithmic ideas and analyses that align with the intentions and limitations of practitioners in real world settings.

As pure exploration research has matured there has been a recent focus on problem settings where can design algorithms that adapt to the hardness of a given problem instance. Moving beyond a worst-case analysis of algorithms the goal is now to understand when it is possible, and how to design, algorithms which can quickly make selections on easy problem instances while still guarding against failure on harder instances. In this thesis we explore problem formulations from both the fixed confidence and fixed budget perspective, discussing theoretical and algorithmic advancements in each, some of which are novel, as well as the relationships between different problem settings. We would be remiss to not mention that a highly active area of pure exploration research is the extension of theory and algorithms toward bandit problems with structured arms such as the immensely practical linear bandit setting (Abe and Long,

1999; Degenne, Menard, et al., 2020; Yinglun Zhu et al., 2022). This line of research looks to design algorithms which make use of the shared information between the arms of an instance in order to learn optimal recommendations faster than is possible in a more general unstructured setting. In this thesis we will restrict our focus to unstructured pure exploration problems in multi-armed bandits, with a brief foray into online convex optimization. Some other topics in pure exploration we will not cover include: 'combinatorial pure exploration' (for example 'Top-K' or 'multiple identification' problems) in which the goal is to return a *set* of arms at the end of sampling (Séebastian Bubeck et al., 2013; Kalyanakrishnan et al., 2012) or identify (an) arm(s) with mean(s) above a given threshold (S. Chen et al., 2014), and the infinitely-armed bandit setting where the goal is to design algorithms with performance guarantees when the number of arms is larger than the number of samples the algorithm can take. Success in this setting relies on making some structural or distributional assumption on the means of the arms on instances (Aziz et al., 2018; Carpentier and Valko, 2015; de Heide et al., 2021).

1.4 Contributions

This thesis covers a comprehensive set of results in core problem settings for pure exploration in multi-armed bandits. In doing so we present previous results from the literature in addition to several original contributions. The main novel contributions of this thesis are as follows:

- We show an instance-dependent sample complexity lower bound for online stochastic convex optimization (Theorem 3.16) which complements a related result by Yuancheng Zhu et al. (2016). The value of this lower bound is that it provides an alternative instance-dependent characterization of problem hardness from the 'local-minimax' framework studied previously.

- We prove a worst-case lower bound on the simple regret of algorithms which use uniform sampling (Theorem 4.10). Previous bounds (Lattimore and Szepesvári, 2020, Exercise 33.2) held only for algorithms which recommend the empirically best arm after sampling has concluded whereas the new bound holds for any recommendation scheme which is equivariant under permutations of the arm indices of the instance.

- We provide a worst-case characterization of the trade-off present in minimizing simple

regret versus best arm identification error with oracle-allocation strategies (Section 4.2.1). By focusing on a simple problem setting and considering an appropriate scaling of the sampling budget with the size of the instance we are able to analyze a problem instance for which obtaining rate-optimal losses on both objectives simultaneously is not possible.

- We discuss new analyses[2] for the fixed budget ε-good arm identification setting (Section 4.3). By tightly characterizing the probability that uniform sampling returns sets of high quality arms we show that an existing algorithm has a nearly-optimal sample complexity for identifying an ε-optimal arm with a fixed budget on certain gap structures. This result closes upper and lower bounds in the related $(n, m, \varepsilon, \delta)$-PAC setting studied by Chaudhuri and Kalyanakrishnan (2017).

[2]These results are based on research conducted with collaborators at the University of Arizona. See the preface of this thesis for more details.

Chapter 2

Background

The multi-armed bandit setting extends the traditional sequential statistics setting to allow the statistician to make use of previous observations to decide the sequence of distributions that they sample from. We define an n-armed bandit instance ν as a finite collection of distributions $(\nu_i)_{i=1}^n$ over $\mathbb{R}$ with respective means $(\mu_i(\nu))_{i=1}^n$. We will typically shorthand $\mu_i \coloneqq \mu_i(\nu)$ when there is no risk of confusion. Unless specified otherwise we will restrict our attention to the class of 1-subgaussian bandits – i.e. instances where the rewards from arm $i \in [n]$ are 1-subgaussian: $\forall \lambda \in \mathbb{R},\ \mathbb{E}_{X \sim \nu_i}\left[\exp(\lambda(X - \mu_i)\right] \leq \exp\left(\lambda^2/2\right)$, where we deviate from the classical definition to include random variables with non-zero mean.[1] We will also make the assumption that all arm means lie in the $[0, 1]$ interval, where this second assumption is made to improve the clarity of various proofs throughout, at the cost of losing some generality.

An algorithm (also referred to as a learner, player, agent or a strategy) interacts with ν over a number of rounds, either deterministic and fixed in advance in the fixed budget problem setting, or determined by the algorithm's own (possibly random) stopping time in the fixed confidence setting. In the t^{th} round of interaction the player selects an arm index $A_t \in [n]$ according to some distribution $\pi_t(A_1, X_1, \ldots, A_{t-1}, X_{t-1})$ and observes a reward $X_t \sim \nu_{A_t}$. This interaction between a learner with data collection policy $\pi = (\pi_t)_{t \geq 1}$ and a bandit instance ν allows us to define the filtration $\mathcal{F} = (\mathcal{F}_t)_{t \geq 1}$, $\mathcal{F}_t \coloneqq \sigma\left(A_1, X_1, \ldots, A_t, X_t\right)$ and induces a probability measure $\mathbb{P}_{\nu,\pi}$ over possible interaction histories. After the interaction phase has finished we enter the selection phase where the learner makes a (possibly random) selection

[1] Intuitively, assuming subgaussianity corresponds to considering bandit arms with tails that are no heavier than that of a gaussian distribution, which in turn means that we expect the sample means of such arms to concentrate at least as quickly as those of from a gaussian distribution. This class includes distributions with sufficiently bounded support, as well as gaussian distributions with variance less than one.

according to $\psi_\tau : \mathcal{F}_\tau \to [n]$, where ψ_τ is a $\mathcal{F}_\tau$ measurable function and τ is either a stopping time with respect to $\mathcal{F}$[2] defined by the learner in the fixed confidence setting, or a deterministic budget constraint $\tau \coloneqq T$ in the fixed budget case. We have defined our probability space around a sequential outcome model, however equivalent models exist that produce identical probability measures over interaction histories. An example that we will use later is a reward table model where we consider a (possibly infinite) table of rewards $X_{t,i}$ where for each $i \in [n]$, $X_{t,i}$ is drawn according to ν_i for all t. In this model we think of the interaction as the learner making its way through each row of the pre-drawn table, with the t^{th} row corresponding to the t^{th} round of interaction and where the reward observed by the learner in round t is the value at column A_t, X_{t,A_t}.

We will use the arm indexing convention $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$ unless specified otherwise, with the understanding that we retain full generality since we typically assume that algorithms do not depend explicitly on the arm indices – in the real world a learner usually has to assume that the indexing of arms is completely arbitrary. As well, for the purposes of clearer exposition in proofs and theorems we will usually omit the use of rounding symbols such as $\lceil \cdot \rceil$ since the results we are interested in are primarily those that concern the order-wise growth of quantities with respect to problem and algorithm specific parameters that are unaffected by rounding details.

2.1 The Objective(s)

At a high level the goal of pure exploration is to design algorithms that efficiently gather data that can be used to make a high quality final selection. As we will discuss in later chapters there is no one clear objective for pure exploration. The notion of a 'high quality final selection' varies based on the reason for running the experiment, as well as the limitations of the available data. For example the New Yorker Caption Contest (Tanczos et al., 2017) is an ongoing competition in which crowd-sourced caption suggestions are gathered for single-panel cartoons. Fig. 2.1 explains the process in more detail. When designing an algorithm to collect voting information for the contest it would be reasonable to design an algorithm that on average returns a really funny caption (maximizes $\mathbb{E}_{\nu,\pi}[\mu_{\psi_\tau}]$), but in the spirit of

[2] i.e. a random variable τ taking values in $\mathbb{N} \cup \{\infty\}$ with the property that $\mathbb{1}\{\tau \leq t\}$ is $\mathcal{F}_t$-measurable for $t \geq 1$.

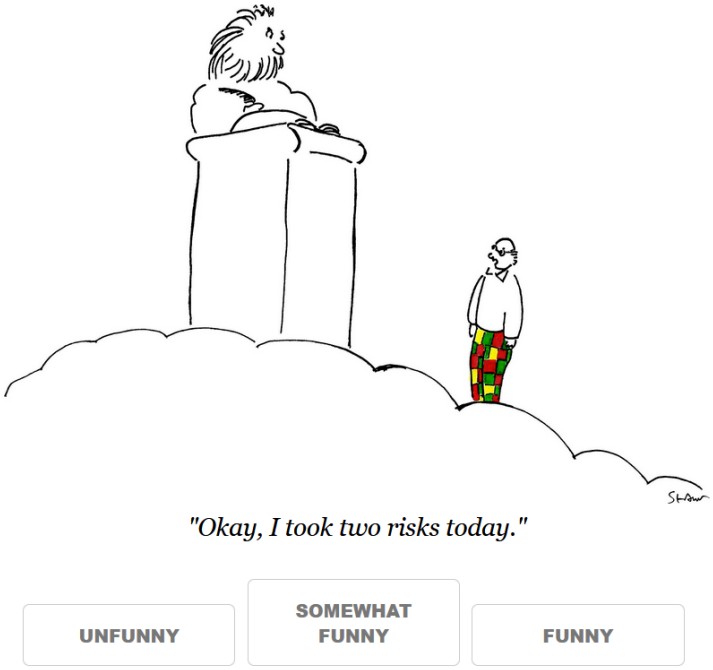

Figure 2.1: A sample voting selection for the New Yorker caption contest (`https://www.newyorker.com/cartoons/vote`) on the week of September 25th 2022. Users vote on the quality of the caption served to the user by the website. Given that each user will have their own preferences over captions we can treat an assessment of caption quality from a given user as a noisy sample of the true population appraisal of the caption.

awarding the true winner of the competition we may alternatively choose to maximize the probability that the recommended caption is truly the submission with the highest score from the population (maximizes $\mathbb{P}_{\nu,\pi}\left(\psi_\tau \in \arg\max_{i\in[n]} \mu_i\right)$). On the other hand consider a genomics experiment where the numbers of arms (genes) may be comparable to the sample size constraints (participants in the experiment). If the goal is to select a candidate genome for further study given limited samples then the experimenter's primary concern may not be resolving which of two similar options is best, and it may instead be more suitable to design an algorithm that seeks to ensure that the returned candidate is of high quality as quickly as possible. This idea can be characterized in several ways and will come up when we discuss algorithms that satisfy a 'probably approximately correct' (PAC) criteria, or minimize some statistic of the 'simple regret', the difference between the largest true mean of the arms being considered, and the true mean of the arm that was selected. The differences between these forms of objectives are a central consideration in this thesis and will be returned to in the remaining chapters.

2.2 Additional Notation

We have included a collection of some of the notation that will appear throughout the later chapters.

- For a bandit instance with means $(\mu_i)_{i=1}^n$ sorted in non-decreasing order as specified previously we define the suboptimality gap of the i^{th} arm as $\Delta_i \coloneqq \mu_1 - \mu_i$ for $2 \leq i \leq n$, and $\Delta_1 \coloneqq \Delta_2$ by convention.

- We use the term ε-*good* to refer to an arm for which $\Delta_i < \varepsilon$ for some $\varepsilon > 0$, i.e. an arm with a mean value within ε of the optimal value. Correspondingly we use ε-*bad* to refer to an arm for which $\Delta_i \geq \varepsilon$, i.e. an arm that is suboptimal by at least ε. For a fixed value of $\varepsilon > 0$ any given arm is one of either ε-good or ε-bad.

- We denote the number of times that a sample has been observed from arm $i \in [n]$ up to round t, $N_i(t) \coloneqq \sum_{s=1}^{t-1} \mathbb{1}\{A_s = i\}$ and use $\hat{\mu}_i(t) \coloneqq N_i(t)^{-1} \cdot \sum_{s=1}^{t-1} \mathbb{1}\{A_s = i\} X_i$ to refer to the sample mean of arm i of samples collected from arm i up to round t – note that this differs from another common definition of $\hat{\mu}_i(t)$ as the sample mean of arm i after t samples from *that arm*.

- We use the shorthand $[n] \coloneqq \{1, 2, \ldots, n\}$ for $n \in \mathbb{N}$, as well as $[n : m] \coloneqq \{n, n+1, \ldots, m\}$ for $n < m \in \mathbb{N}$.

- For a set A in universe Ω we will use an over-bar $\overline{A}$ to denote the complement of A in Ω, $\overline{A} \coloneqq \Omega \setminus A$.

- $\log$ refers to the natural logarithm unless another base is explicitly specified, e.g. $\log_2$.

- For a finite set A we denote the uniform distribution over elements in A in an implicit fashion: $\mathbb{P}_{X \sim A}(X = x) \coloneqq \mathbb{1}\{x \in A\} \cdot |A|^{-1}$.

- We use $\binom{n}{m}$ to refer to the collection of sets $\{S \subseteq [n] : |S| = m\}$.

- We use the term 'shifted standard normal' to refer to unit variance gaussian distributions with arbitrary means.

2.3 Cumulative Regret Minimization versus Pure Exploration

The usual bandit problem is formulated with the goal of maximizing the cumulative sum of observed rewards, or more specifically when sampling according to π on instance ν with means $(\mu_i)_{i=1}^n$ over T rounds, $\mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^T X_t\right]$ (Auer et al., 2002). This is usually framed as the equivalent problem of minimizing the cumulative regret, $R_T(\nu,\pi) \coloneqq T \cdot \mu_1 - \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^T X_t\right]$. A more illuminating form of the regret is established via the regret decomposition lemma (Lattimore and Szepesvári, 2020, Lemma 4.5):

$$R_T = \sum_{i \in [n]} \Delta_i \cdot \mathbb{E}\left[N_i(T)\right],$$

where we drop the notational dependence on π, ν where there is no ambiguity. Section 2.3 illustrates that in order to have cumulative regret that scales sublinearly with T an algorithm has to sample suboptimal arms sublinear times in expectation. It follows that algorithms that achieve sublinear cumulative regret necessarily eventually play the optimal arm most often and are therefore a reasonable first approach for solving pure exploration problems. On the other hand, as pointed out by Sébastien Bubeck, Rémi Munos, and Stoltz (2009) algorithms that achieve optimal cumulative regret offer worse long-term pure exploration performance than even non-adaptive algorithms when the sampling budget becomes large. This is due to their need to sparingly sample suboptimal arms after the initial rounds of interaction. This aggressively directed sampling prevents these algorithms from collecting enough information about suboptimal arms and limits their ability to make an optimal recommendation at the end of the interaction phase. Whereas cumulative regret minimization in bandits is a paradigm designed to explore the optimal trade-off between exploration and exploitation, pure exploration is solely concerned with collecting decisive information about arms as efficiently as possible. With that said we will see that many of the algorithmic ideas from the extensive literature on the cumulative regret setting can be adapted to construct efficient algorithms for the pure exploration setting.

2.4 Finite-Time Concentration Inequalities

The majority of the results and analysis discussed in this thesis are non-asymptotic in nature. In order to provide probabilistic guarantees on the performance of algorithms interacting with stochastic reward sequences we rely on some common non-parametric distributional assumptions for the distributions of the rewards from arms, namely that they are bounded or subgaussian random variables.

Our interest in subgaussian variables is that they offer encompass many reasonable models and allow for finite-time concentration guarantees in line with those of gaussian random variables. We list some useful properties of bounded and or subgaussian random variables without proof. More details can be found in texts that cover elements of non-asymptotic statistical theory such as Wainwright (2019).

Theorem 2.1 (Hoeffding's inequality for σ-subgaussian random variables)**.** *Let X be a zero-mean independent σ-subgaussian random variable for some $\sigma > 0$. Then*

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

Making use of the properties of the moment generating function of sums of independent random variables we arrive at the following corollary that will see extensive use in later chapters:

Corollary 2.1.1. *Let $X_1, X_2, \ldots, X_n$ be zero-mean independent σ-subgaussian random variables for some $\sigma > 0$. Then*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq n \cdot \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

Comparing Theorem 2.1 to an upper bound for the gaussian complementary distribution function:

$$\int_{\varepsilon}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \leq \int_{\varepsilon}^{\infty} \frac{x}{\varepsilon} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx = \frac{1}{\sqrt{2\pi\varepsilon^2}} \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

we see that for values for ε large enough for the effect of the exponential decay to dominate the factor of $1/\varepsilon$ the Hoeffding bound is tight up to constant factors.

As promised we can apply results for subgaussian random variables to bounded random variables:

Theorem 2.2. *Let X be a bounded random variable $X \in [a, b]$ almost surely, then X is $(b-a)/2$-subgaussian.*

We will focus on 1-subgaussian instances for the remainder of this thesis, however we can see from Theorem 2.1 that in a more general scenario with a common finite variance $\sigma^2 > 0$, a simple re-scaling of $X \to X/\sigma$ takes us back to the 1-subgaussian case.

2.5 Information-Theoretic Lower Bounds

In the next few chapters we will look at various measures of the performance of pure exploration algorithms such as the probability of bad outcomes occurring or looking at the distribution of the number of samples an algorithm takes before it selects an arm. Our interest in lower bounds is twofold. Constructing lower bounds can help us understand the fundamental statistical challenges that underlie the problems that we are interested in, which can sometimes lead to improved algorithms. As well, the result of these bounds can help determine whether or not there is room to meaningfully improve on the performance of the algorithms we are considering or whether they have run up against unavoidable statistical limits and are in this sense optimal. Finding a meaningful foundation for the notion of optimality in problems of a statistical nature is challenging and requires carefully chosen restrictions on both the problem class being considered as well as the form of the solutions we allow.

Consider a needle-in-a-haystack problem where we are tasked with selecting the arm with the highest mean on an n-armed, shifted standard normal bandit instance with one arm that has mean 1 and all other arms having mean 0 – but we are not allowed to take any samples from them beforehand. Clearly there is no room for clever algorithmic design here. We want a mathematical framework that is concrete enough that it allows us to write theorems that reflect our understanding of the hardness of this problem while still being general enough that we can analyse the situations that we are interested in.

The haystack problem can be understood with a basic algorithmic correctness argument: if the arms of the bandit instance are shuffled randomly prior to the game then clearly no algorithm can do better than selecting the correct arm with probability $1/n$ on average. More concretely, assuming that the arm indices of the instance ν are permuted by $\sigma : [n]^n \to [n]^n$ drawn uniformly at random from the permutation group Σ^n on sets of size n. Then for any

algorithm Alg

$$\mathbb{E}_{\sigma\sim\Sigma^n}\left[\mathbb{P}_{\sigma(\nu),\mathsf{Alg}}\left(\psi=\sigma(1)\right)\right]=1/n,$$

where we overload our notation to let $\sigma(\nu)$ denote the permuted bandit instance and $\sigma(1)$ be the index in $[n]$ that 1 is mapped to by σ.

Proof. Fix some algorithm Alg, allowing randomization. Given that the algorithm never observes any samples from any arms the probability of choosing a given action is independent of the instance and we have that for any ν, $\mathbb{P}_{\nu,\mathsf{Alg}}\left(\psi=i\right)=p_i$ for some $p\in\mathcal{M}_n$, the probability simplex in dimension n. It follows that

$$\begin{aligned}
\mathbb{E}_{\sigma\sim\Sigma^n}\left[\mathbb{P}_{\sigma(\nu),\mathsf{Alg}}\left(\psi=\sigma(1)\right)\right] &= \frac{1}{|\Sigma^n|}\sum_{\sigma\in\Sigma^n}\mathbb{P}_{\sigma(\nu),\mathsf{Alg}}\left(\psi=\sigma(1)\right)\\
&= \frac{1}{n!}\sum_{i=1}^{n}\left(\sum_{\sigma\in\Sigma^n:\ \sigma(1)=i}\mathbb{P}_{\sigma(\nu),\mathsf{Alg}}\left(\psi=i\right)\right)\\
&= \frac{(n-1)!}{n!}\sum_{i=1}^{n}p_i\\
&= 1/n.
\end{aligned}$$

□

Building on this idea we can consider a more realistic problem where an algorithm has to make a decision as to which of the n shifted standard normal arms is optimal after some $T\geq n$ samples. Our previous argument no longer holds since the algorithms can make decisions based on interactions with the instance. We still want to carry our intuition from the previous, simpler case in that given the limited sampling budget T there is a limit to the statistical extent any learner can distinguish which permutation of the instance it is interacting with based on noisy samples. Namely there is a sense in which the distribution of rewards we expect from instances ν and $\sigma(\nu)$ with different optimal arms are similar, and therefore the distribution over the arms recommended by any learner will be similar on either instance, incurring errors on at least one of the instances with some probability. Theorem 4.2 in Chapter 4 is such a result, and makes use of tools introduced below.

2.5.1 Information Inequalities

Information theory is a tool that allows us to translate the intuition that if probability distributions P and Q are close in some sense then we should be able to show that for some

event $\mathcal{E}$ we are interested in, $|P(\mathcal{E}) - Q(\mathcal{E})|$ is small. Going back to the previous example: suppose we have two probability measures that are induced by the interactions of an algorithm on two different bandit instances which are identical to one-other than a small difference in the mean of one arm. Then we should be able to show that the probability that the algorithm returns arm 1 on the first instance is close to the probability that the algorithm returns arm 1 on the second instance. In the chapters that follow we will rely on a handful of useful results from information theory that allow us to flesh out this intuition more formally.

We assume that the reader is familiar with the Kullback-Liebler (KL) divergence: for two probability distributions P and Q over a common measurable space $(\mathcal{X}, \mathcal{G})$ such that $P \ll Q$ we define

$$\mathrm{KL}\left(P, Q\right) = \int p \log\left(\frac{p}{q}\right) d\lambda \tag{2.1}$$

where $\lambda = P + Q$ and $p = \frac{dP}{d\lambda}$, $q = \frac{dQ}{d\lambda}$ are the respective Radon-Nikodym derivatives with respect to λ.

The KL-divergence has many convenient properties as a divergence, for instance the KL-divergence is strictly positive unless $P = Q$, for which the KL-divergence is zero. In this work we will only need to know how to evaluate the KL-divergence between pairs of Normal distributions with common variance, or between Bernoulli random variables.

Lemma 2.3.

$$\mathrm{KL}\left(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)\right) = \frac{\left(\mu_1 - \mu_2\right)^2}{2\sigma^2},$$

where $\mathcal{N}(\mu, \sigma)$ denotes a normal distribution over $\mathbb{R}$ with mean $\mu \in \mathbb{R}$ and variance $\sigma > 0$, and

$$\mathrm{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q},$$

where $\mathcal{B}(p)$ denotes a Bernoulli distribution over $\{0, 1\}$ with $\mathbb{P}_{X \sim \mathcal{B}(p)}\left(X = 1\right) = p$.

The KL divergence plays an important role as a choice of measure of 'closeness' that we were referencing earlier, allowing us to bound the difference in the probability of an event under two probability measures as a function of their KL-divergence.

Let P and Q be two probability distributions over a common measurable space $(\mathcal{X}, \mathcal{G})$. The total variation distance between P and Q is defined as $\mathrm{TV}\left(P, Q\right) \coloneqq \sup_{\mathcal{E} \in \mathcal{G}} |P(\mathcal{E}) - Q(\mathcal{E})|$.

The KL-divergence provides a bound on the degree to which the probabilities of events can change between distributions P and Q.

Theorem 2.4 (Bretagnolle-Huber Inequality (Bretagnolle and Huber, 1978))**.**

$$\mathrm{TV}\left(P, Q\right) \leq 1 - \frac{1}{2}\exp\left(-\mathrm{KL}\left(P, Q\right)\right).$$

Another way of framing Theorem 2.4 is to note that this is equivalent to the statement that for any event $\mathcal{E} \in \mathcal{G}$,

$$P(\mathcal{E}) + Q(\bar{\mathcal{E}}) \geq \frac{1}{2}\exp\left(-\mathrm{KL}\left(P, Q\right)\right). \tag{2.2}$$

It will be useful to specialize some of the results from information theory to the distributions induced by learners interacting with bandit instances. In the beginning of this chapter we defined a probability space for stochastic bandits interacting sequentially with an instance. Let $\mathbb{P}_{\nu,\pi}$ be the probability measure over interaction histories and arm selections after T rounds of sampling according to π on instance ν. The following lemma characterizes the KL-divergence of such a measure on different bandit instances via the KL-divergence of the distributions of the arms on each instance.

Lemma 2.5 (Divergence decomposition (Lattimore and Szepesvári, 2020, Lemma 15.1))**.** *Let $\nu = (\nu_i)_{i=1}^n$ and $\tilde{\nu} = (\tilde{\nu}_i)_{i=1}^n$ be n-armed bandit instances and let π be some sampling policy that interacts over T rounds of sampling.*

$$\mathrm{KL}\left(\mathbb{P}_{\nu,\pi}, \mathbb{P}_{\pi,\tilde{\nu}}\right) = \sum_{i=1}^{n} \mathbb{E}_{\nu,\pi}\left[N_i(T)\right] \mathrm{KL}\left(\nu_i, \tilde{\nu}_i\right).$$

We will also need to consider divergences between distributions induced by learners with a potentially random $\mathcal{F}$-measurable stopping time τ.

Lemma 2.6 (Stopping-time divergence decomposition)**.** *Let $\nu = (\nu_i)_{i=1}^n$ and $\tilde{\nu} = (\tilde{\nu}_i)_{i=1}^n$ be n-armed bandit instances and let π be some sampling policy that interacts with the instance until some stopping time τ with respect to $\mathcal{F}$, where we assume that $\mathbb{E}_{\nu,\pi}\left[\tau\right] < +\infty$. Let X be an $\mathcal{F}_\tau$-measurable random element, for example the final selection of a fixed confidence pure exploration algorithm, ψ_τ. Let $\mathbb{P}_{(\pi,\nu)X}$ denote the push-forward measure of $\mathbb{P}_{\pi,\nu}$ under X.*

$$\mathrm{KL}\left(\mathbb{P}_{(\pi,\nu)X}, \mathbb{P}_{(\pi,\tilde{\nu})X}\right) \leq \sum_{i=1}^{n} \mathbb{E}_{\nu,\pi}\left[N_i(\tau)\right] \mathrm{KL}\left(\nu_i, \tilde{\nu}_i\right).$$

The proof of Lemma 2.6 can be found in Section A.1. The crux of the proof is a simple data-processing inequality followed by a careful analysis of Radon-Nikodym derivatives of probability measures on a stopped filtration.

These tools will be used to extend the simple argument at the beginning of this section towards more complex scenarios that allow us to explore the limits of interactive statistical estimation.

2.5.2 Worst-Case versus Instance-Dependent Bounds

In later chapters we will distinguish between worst-case (sometimes called minimax) lower bounds and their more instance-dependent counter parts. In order to distinguish between these forms of lower bounds we need to be clear about the problem we are solving. Worst-case lower bounds are usually simpler to analyze, and are often the first step towards analyzing a new learning setting.

We will adopt a relatively abstract framework in the style of Tsybakov (2009) and Wainwright (2019). Consider an estimation problem where we receive a random variable X drawn from some distribution $\mathbb{P} \in \mathcal{P}$, where $\mathcal{P}$ is equipped with a functional θ that maps distributions in $\mathcal{P}$ to some space $\mathcal{X}$. The estimation problem is to construct an estimate $\hat{\theta}$ of $\theta(\mathbb{P}) = \theta^*$ based on X in such a way that we minimize $\rho(\hat{\theta}, \theta(\mathbb{P}))$ where $\rho : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is a semi-metric[3]. We understand $\hat{\theta} = \hat{\theta}(X)$ to be a random variable taking values in $\mathcal{X}$, and so we can obtain a deterministic measure of the quality of $\hat{\theta}$ by taking an expectation $\mathbb{E}_{\mathbb{P}}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]$.

Our goal is to understand the statistical limits of estimating $\theta(\mathbb{P})$ from X. This requires some nuance: how good of an estimator is the naive estimator that always returns $\tilde{\theta} \in \Theta$ regardless of the value of X? Clearly if $\theta^* = \tilde{\theta}$ then this estimator is optimal in that it minimizes $\mathbb{E}_{\mathbb{P}}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right] = 0$. On the other hand this estimator can be arbitrarily bad for other values of $\theta(\mathbb{P})$. There are several ways to get around this issue. For now we will consider an approach where we allow θ^* to be chosen in an adversarial fashion and evaluate the performance of an estimator $\hat{\theta}$ on its worst-case scenario: $R(\hat{\theta}, \mathcal{P}) = \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]$. We can then define the minimax risk of estimating $\theta(\mathbb{P})$ on the class of distributions $\mathcal{P}$ as

$$R(\mathcal{P}) = \inf_{\hat{\theta}} R(\hat{\theta}, \mathcal{P}), \tag{2.3}$$

[3]Following Wainwright (2019) we define a semi-metric as a function that satisfies the requirements of a metric, other than being positive semi-definite rather than positive definite.

where the infimum over estimators considers all measurable functions $\hat{\theta}$ that map elements of outcome space of X to elements of $\mathcal{X}$. The term 'minimax' comes from the inf-sup form above. Grounding ourselves in a pure exploration setting, X could be a batch of samples observed from taking $B \in \mathbb{N}$ samples from each arm of a bandit instance $\nu = (\nu_i)_{i=1}^n$, inducing a law $\mathbb{P}_\nu$ for X, and we can take $\theta(\mathbb{P}_\nu) = \arg\max_{i \in [n]} \mathbb{E}[\nu_i]$ to be the set of arms with the highest mean. Looking at the form of Eq. (2.3) we can see a potential drawback to this particular measure of the hardness of estimation on the class $\mathcal{P}$. In particular this measure revolves around a worst-case distribution (bandit instance) $\mathbb{P}(\hat{\theta})$ for each θ. Depending on the class $\mathcal{P}$ we are considering we may have reason to suspect that a small collection of the distributions we encounter will be significantly harder than others. In that case we expect that a well designed estimator $\hat{\theta}$ can achieve a lower risk than $R(\mathcal{P})$ on most of the instances we encounter, however such an estimator has a minimax risk that is no smaller than an estimator $\hat{\theta}'$ that achieves a uniformly minimax risk $\mathbb{E}_{\mathbb{P}}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right] = R(\hat{\theta}, \mathcal{P})$ on all distributions $\mathbb{P}$. All else being equal if $\hat{\theta}$ and $\hat{\theta}'$ have the same minimax risk then we should prefer $\hat{\theta}$. To formalize this idea we will switch to a different way of understanding the hardness of $\mathcal{P}$.

Our original motivation for examining the quantity $R(\hat{\theta}, \mathcal{P}) = \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right]$ was to formalize the disadvantage of so called 'pointwise' estimators which are necessarily only good on a small set of instances. Instead we can explicitly restrict our attention to some class of ε-consistent estimators

$$\Theta_{\varepsilon,\mathcal{P}} = \left\{\hat{\theta} : \forall \mathbb{P} \in \mathcal{P},\ \mathbb{E}_{\mathbb{P}}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right] \leq \varepsilon\right\} \tag{2.4}$$

for some $\varepsilon > 0$. This allows us to define an alternative measure of hardness:

$$R(\mathbb{P}; \Theta_{\varepsilon,\mathcal{P}}) = \inf_{\hat{\theta} \in \Theta_{\varepsilon,\mathcal{P}}} \mathbb{E}_{\mathbb{P}}\left[\rho(\hat{\theta}, \theta(\mathbb{P}))\right] \tag{2.5}$$

where this value characterizes the risk of ε-consistent estimators on the class $\mathcal{P}$ for the specific *instance* $\mathbb{P}$. So long as $\Theta_{\varepsilon,\mathcal{P}}$ is non-empty a, reasonable class of estimators that excludes pathological estimators such as point-estimators (e.g. the set of learners that return the correct arm with probability at least $1/2$) then Eq. (2.5) provides a lower bound on the risk that more closely predicts the performance limitations that we expect to see in practice. There are many different ways to define the class of consistent, or well behaved estimators that we may want to restrict our attention to. A common approach in the pure exploration context is to consider

'symmetric' learners that are agnostic to the arm indices of the instance they interact with, i.e. learners that do not explicitly use the indices of the arms to make decisions as opposed to, for example, algorithms that always return arm 1. This is a natural choice since in general we cannot expect any information about the means of the arms to be encoded in the arm indices in practice; for example, arms are unlikely to be sorted by their means.

We will encounter both worst-case as well as more instance-dependent upper and lower bounds in later chapters, but a special emphasis will be placed on instance-dependent results as they are often stronger in the sense that in some cases instance-optimal algorithms can be shown to be minimax optimal, and also since they more faithfully capture the limits of adaptivity in sequential statistical problems. For example we will see that in several cases non-adaptive algorithms with pre-determined sampling procedures have nearly optimal worst-case pure exploration performance, whereas it turns out that algorithms must adapt their sampling behaviour on certain instances in order to approach instance-dependent lower bounds.

Chapter 3

Pure Exploration with Fixed Confidence

In pure exploration with fixed confidence we are tasked with constructing algorithms that require the fewest samples possible to return a selection that satisfies some constraint on the quality of the recommended arm with some fixed probability. For instance: suppose that you are tasked with polling which of several politicians is the most popular among the readership of a website that you work for, and that after the poll is complete you are to announce the favourite. To conduct the poll unique users who come to the website are asked for their feelings (dislike, neutral, like) on a candidate chosen by the polling algorithm. Keeping the poll running for more users will provide you with a better estimate for who the true favourite is but your boss is anxious to get the poll done as quickly as possible. At the same time, you want some assurance that after the completion of your poll you have a high (say, 95%) chance of identifying the correct politician. The algorithms in this section address the problem of running and analyzing this kind of poll with as few samples as possible while still providing high confidence on the results.

We will begin with results for the PAC setting, however the revival of developments in pure exploration in the mid 2010's was centred on the study of the best-arm-identification problem that we will explore afterwards.

3.1 PAC Identification of Good Arms

The probably approximately correct (PAC) problem setting is one of the oldest and simplest scenarios for studying pure exploration in bandits. The goal is to design (ε, δ)-PAC algorithms, i.e. algorithms $\mathsf{Alg} = (\pi, \tau, \psi)$ for which $\mathbb{E}_{\pi,\nu}[\tau] < \infty$ and $\mathbb{P}_{\nu,\pi}(\Delta_{\psi_\tau} > \varepsilon) \leq \delta$ on any

Definition 3.1. Pure Exploration with Fixed Confidence

Require: Confidence level δ, stopping rule τ, loss function $\ell : [n] \to \mathbb{R}^+$

while $t \leq \tau$ **do**
    Select arm $A_t \sim \pi_t(\mathcal{F}_{t-1})$
    Observe reward $X_t \sim \nu_{A_t}$
    (Optionally) Output $\mathcal{F}_t$-measurable arm selection ψ_t
end while
Receive loss $\ell(\psi_\tau, \nu)$

1-subgaussian instance ν with the fewest samples possible, where again $\Delta_i = \mu_1 - \mu_i$ is the suboptimality gap of arm $i \geq 2$ and τ is a stopping rule with defines an almost-surely finite $\mathcal{F}$-measurable stopping time. This problem has been well studied; in the early 2000s Even-Dar et al. (2002) introduced Median Elimination, an algorithm with a sample complexity that matches a worst-case lower-bound up to a constant factor (Mannor and Tsitsiklis, 2003).

This settings is relatively straightforward due to its parameterized form: as we will see (ε, δ)-PAC algorithms only need to adapt to the the hardness (or easiness) of the problem instance they interact with up to a tolerance of of $\varepsilon > 0$ at a level δ. This simplifies the worst-case design and analysis of these algorithms, but also limits the the practical significance of this problem setting. When deploying these algorithms in the world experimenters may not have fore-knowledge of the appropriate choice of ε for their specific application. Some algorithms for the PAC setting have sampling policies π that depend explicitly on the choice of ε, and so reducing the value of ε mid-experiment may destroy any correctness guarantees for these algorithms. In Section 3.1.2 we will discuss alternate problem settings that attempt to address some of these issues.

We will begin by first constructing a non-adaptive algorithm that comes within a logarithmic factor of the minimax sample complexity for the (ε, δ)-PAC setting.

Theorem 3.1. *Fix $\varepsilon > 0, \delta \in (0, 1)$. Let ν be an n-armed 1-subgaussian bandit instance. Taking $\frac{8}{\varepsilon^2}\log(\frac{2n}{\delta})$ samples from each arm of the arms in ν and recommending the empirically best arm $\psi_\tau = \arg\max_{i \in [n]} \hat{\mu}(\tau)_i$ is an (ε, δ)-PAC algorithm with deterministic sample complexity $\tau = \frac{8n}{\varepsilon^2}\log(\frac{2n}{\delta})$.*

Proof. Making use of the 1-subgaussian assumption on arms we can construct Hoeffding confidence intervals for all arms. Let $\mathcal{E} = \{i \in [n] : |\hat{\mu}_i(T) - \mu_i| < \varepsilon/2\}$ be an event on which

the empirical means of all arms remain $\varepsilon/2$ close to their true means. This event occurs with probability at least $1-\delta$. To see this we can use the independence of the arm means to claim that

$$\mathbb{P}_{\nu,\pi}\left(\bar{\mathcal{E}}\right) = \sum_{i=1}^{n} \mathbb{P}_{\nu,\pi}\left(|\hat{\mu}_i(\tau) - \mu_i| \geq \varepsilon/2\right).$$

Then for each $i \in [n]$ since $N_i(\tau) = \frac{8}{\varepsilon^2}\log(\frac{2n}{\delta})$ we can use a Hoeffding bound (Corollary 2.1.1) to obtain $\mathbb{P}\nu, \pi\left(|\hat{\mu}_i(\tau) - \mu_i| \geq \varepsilon/2\right) \leq 2\exp\left(\frac{-\varepsilon^2 N_i(\tau)}{8}\right) = \delta/n$. It follows that on $\mathcal{E}$, if we select the arm with the empirical best mean $\psi_\tau = \arg\max_{i\in[n]} \hat{\mu}_i$, then

$$\Delta_{\psi_\tau} = \mu_1 - \mu_{\psi_\tau} \leq (\hat{\mu}_1 + \varepsilon/2) - (\hat{\mu}_{\psi_\tau} - \varepsilon/2) \leq \varepsilon.$$

□

This uniform exploration approach has a (deterministic) sample complexity of $\tau = \frac{8n}{\varepsilon^2}\log\left(\frac{2n}{\delta}\right)$. One of the contributions of Even-Dar et al. (2002) was to show that with a new algorithm the union bound term $\log(n/\delta)$ could be replaced with $\log(1/\delta)$, and a lower bound from Mannor and Tsitsiklis (2003) proved that this later result was optimal in a minimax sense.[1]

3.1.1 Median Elimination

The key to avoiding the use of $\log(n/\delta)$ samples per arm lies in the following lemma, which is inspired by Even-Dar et al. (2002, Lemma 1):

Lemma 3.2. *Consider an arbitrary 1-subgaussian bandit instance ν with suboptimality gaps $(\Delta_i)_{i=2}^n$. Draw $8\log(3\delta)/\varepsilon^2$ samples uniformly from each arm and select the top half of arms according to their empirical means, $\hat{\mathcal{S}} \in [n]^{n/2}$. Then*

$$\mathbb{P}_\nu\left(\max_{i\in\hat{\mathcal{S}}} \Delta_i \geq \varepsilon\right) \leq \delta.$$

Proof. We want to upper bound the probability that the top performing half of arms does not contain an ε-good arm. This can only happen if each of the arms in the top half is ε-bad and thus arm 1 is in the bottom half of arms. If we draw $8\log(3\delta)/\varepsilon^2$ samples uniformly from each

[1] One might wonder whether the factor of $\log(n/\delta)$ is an artifact of our analysis of uniform exploration. A similar proof to Theorem 4.10 in Chapter 4 shows that this dependence on $\log(n)$ is unavoidable for any symmetric algorithm (formalized in the next) that uses uniform exploration.

arm then apply a Hoeffding bound we find that the probability that the best arm lies outside of a $\varepsilon/2$ interval around its true mean is less than $1/(3\delta)$. On the event that the best arm is contained within this interval, then in order for the entire top half of arms to not contain an ε-good arm we must have at least $n/2$ ε-bad arms with empirical mean at least $\mu_1 - \varepsilon/2$. By Markov's inequality we have $\mathbb{P}(N_{\varepsilon^+} \geq n/2) \leq \frac{\mathbb{E}\left[N_{\varepsilon^+}\right]}{n/2} \leq \frac{2\delta}{3}$, where $N_{\varepsilon^+} \coloneqq \sum_{i:\Delta_i > \varepsilon} \mathbb{1}\{\hat{\mu}_i \geq \mu_1 - \varepsilon/2\}$. By a union bound we have $\mathbb{P}\left(|\hat{\mu}_1 - \mu_1| \geq \varepsilon/2, N_{\varepsilon^+} \geq n/2\right) \leq \delta$. □

Lemma 3.2 tells us that by returning a fixed *proportion* of the original arms we can control our error probability by balancing the probability that *any* arm lies outside of its confidence interval with the requirement that a similar bad event needs to occur independently and simultaneously on a constant *proportion* of arms in order for a mistake to occur. This allows us to control the mistake probability with $O\left(\log(1/\delta)\right)$ samples per-arm rather than $O\left(\log(n/\delta)\right)$.

Let $\hat{\mu}_i^\ell$ denote the empirical mean of samples taken from arm i in phase ℓ.
Require: Confidence level $\delta \in (0, 1)$, error tolerance $\varepsilon > 0$
1: Set $\mathcal{S}_1 = [n]$
2: Set $\varepsilon_1 = \varepsilon/4$, $\delta_1 = \delta/2$, $\ell = 1$
3: **for** $\ell \in \{1, 2, \ldots, \log_2(n)\}$ **do**
4: Sample each arm $i \in \mathcal{S}_\ell$ $\frac{1}{(\varepsilon_\ell/2)^2}\log(3/\delta_\ell)$ times, record sample means $\hat{\mu}_i^\ell$
5: $\mathcal{S}_{\ell+1} = S_\ell \setminus \left\{i : \hat{\mu}_i^\ell < \hat{\mu}_{[n/2^\ell]}^\ell\right\}$, the set of arms that perform as well as the median
6: $\varepsilon_{\ell+1} = \frac{3}{4}\varepsilon_\ell;\ \delta_{\ell+1} = \delta_\ell/2$
7: **end for**
8: Select $\psi_t = \arg\max_{i \in [n]} N_i(t)$

Algorithm 1: Median Elimination (Even-Dar et al., 2002)

Median Elimination operates in $\log_2 n$ phases, eliminating the worst half of the remaining arms in each round and allocating an increasingly number of samples per-arm in later rounds.

Theorem 3.3 (Even-Dar et al. (2002, Theorem 4))**.** *Fix $\varepsilon > 0, \delta \in (0, 1)$. Algorithm 1 is (ε, δ)-PAC and uses fewer than $O\left(n \cdot \varepsilon^{-2}\log(1/\delta)\right)$ samples where O hides constant factors.*

As mentioned earlier the heart of the analysis of Algorithm 1 is contained in Lemma 3.2. The second innovation of Median Elimination is the use of a geometrically decreasing schedule of the ε_ℓ and δ_ℓ used to set the arm-sampling budget over rounds $\ell \leq \log_2(n)$. In each round the bottom half of arms are eliminated until the one arm remains. Repeatedly applying Lemma 3.2 as well as the use of the geometric sequence of $\varepsilon_\ell, \delta_\ell$ ensure that the cumulative increase in the

sub-optimality of the best remaining arm in each phase is bounded by ε with probability at least $1-\delta$ using only $O(\frac{n}{\varepsilon^2}\log(1/\delta))$ total samples. To see how this works note that if Lemma 3.2 holds for each round then the cumulative suboptimality gap of the returned arm is bounded by

$$\sum_{\ell=1}^{\log_2 n} \varepsilon_\ell = \frac{\varepsilon}{4}\sum_{\ell=0}^{\log_2 n-1}\left(\frac{3}{4}\right)^\ell = \frac{\varepsilon}{4}\left(4(1-(3/4)^{\log_2 n})\right) \leq \varepsilon.$$

A similar geometric series over the δ_ℓ provides the bound on the total error probability and an arithmetico-geometric series allows us to bound the total number of samples taken by the algorithm. One can verify that if we had used fixed values for $\varepsilon_\ell = \varepsilon_0, \delta = \delta_0$ instead, the sample complexity would necessarily be of order $\tau = \Omega\left(\frac{n\log^2(n)}{\varepsilon^2}\log\left(\log(n)/\delta\right)\right)$. We will see that Median Elimination matches the worst-case complexity for this problem setting, however due to its deterministic (and therefore necessarily worst-case) sampling budget it is incapable of achieving the instance-optimal sample complexity on easier problem instances.

3.1.2 The Sample Complexity of PAC Identification of Good Arms

Theorem 3.4 (Mannor and Tsitsiklis (2003, Theorem 1))**.** *Fix $\varepsilon \in (0, 1/4), \delta \in (0, e^{-4}/4)$ and $n \geq 2$. For any (ε,δ)-correct algorithm* Alg *there exists an n-armed Bernoulli bandit instance ν such that*

$$\mathbb{E}_{\nu,\mathsf{Alg}}[\tau] \geq C_1\frac{n}{\varepsilon^2}\log\left(\frac{C_2}{\delta}\right),$$

where $C_1, C_2 > 0$ are universal constants.

The proof of Theorem 3.4 is typical of pure-exploration lower bounds in that it actually builds off of a stronger statement: Theorem 3.4 holds even when Alg knows the means of the arms up to a permutation, and has a 'needle-in-a-haystack' structure in that it uses permutations of an instance in the 'slippage'-configuration, i.e. a single optimal arm with all other having a uniform gap of ε.[2]

Theorem 3.4 is worst-case in nature in that it only holds for some instance. An instance-dependent result was provided in the same paper and was later strengthened by Kaufmann et al. (2016):

[2] A similar result can be shown for gaussian instances, however the proof technique used by this and many other works relies on bounding a likelihood ratio and is conceptually simpler for arm distributions with a discrete and finite alphabet.

Theorem 3.5 (Kaufmann et al. (2016, Remark 5))**.** *Fix $\varepsilon > 0, \delta \leq .15$. For any (ε, δ)-PAC algorithm $\mathsf{Alg} = (\pi, \tau, \psi)$ and Bernoulli bandit instance ν with mean vector $(\mu_i)_{i=1}^n$*

$$\mathbb{E}_{\nu,\pi}[\tau] \geq \left[\frac{|\{i : \Delta_i < \varepsilon\}| - 1}{\mathrm{KL}\left(\mathcal{B}(\mu_1), \mathcal{B}(\mu_1 - \varepsilon)\right)} + \sum_{i:\Delta_i \geq \varepsilon} \frac{1}{\mathrm{KL}\left(\mathcal{B}(\mu_i), \mathcal{B}(\mu_1 + \varepsilon)\right)} \right] \log\left(\frac{1}{2.4\delta}\right).$$

By the properties of the KL-divergence of Bernoulli random variables we see that this bound suggests an instance-dependent lower bound that scales like $\sum_{i=2}^k \left(\Delta_i^\varepsilon\right)^{-2} \log(1/\delta)$ for small $\varepsilon > 0$ where $\Delta_i^\varepsilon := \max\{\varepsilon, \Delta_i\}$. This bound was matched up to logarithmic factors by the Exponential-Gap Elimination algorithm (Karnin et al., 2013) which uses Median Elimination as a subroutine and has a sample complexity guarantee

$$\mathbb{E}_{\nu,\pi}[\tau] = O\left(\sum_{i=2}^n \frac{1}{\left(\Delta_i^\varepsilon\right)^2} \log\left(\frac{1}{\delta} \log \frac{1}{\Delta_i^\varepsilon}\right)\right).$$

3.2 Best Arm Identification

Best Arm Identification (BAI) in the fixed-confidence setting is a specialization of the (ε, δ)-PAC problem to $\varepsilon = 0$, sometimes referred to as the δ-correct setting. At first glance the scaling with $\varepsilon > 0$ in Theorem 3.4 would seem to indicate that this problem is futile, and indeed if we consider an instance in the slippage configuration, in the limit that the gap of the $n - 1$ suboptimal arms goes toward zero then it is completely sensible that in the worst case the sample complexity of BAI grows without bound. The interest in the BAI setting comes from the challenge of designing algorithms that adapt to the hardness of the problem instance that they interact with. In the (ε, δ)-PAC setting we have seen that a union bound factor is all that separates non-adaptive algorithms from the optimal worst-case sample complexity. This is not entirely surprising given that the explicitly parameterized problem setting naturally prescribes the smallest widths of the confidence intervals required to separate feasible and infeasible solutions with high probability. In the BAI setting adapting to the suboptimality gaps $(\Delta_i)_{i=2}^n$ is all important. Recent work (e.g. Degenne and Koolen (2019)) has begun to address the setting where there may be more than one optimal arm but we will focus on the classical setting where each instance has a unique optimal arm.

Once again we will begin by analyzing a simple algorithm to better understand the challenges that more efficient BAI algorithms must overcome. Consider designing a uniform exploration (UE) algorithm in this setting. In the general PAC setting we were able to make

explicit use of $\varepsilon > 0$ to decide how many samples to take. In the BAI problem we have no such luck[3]. Instead we will need to make use of a slightly more clever algorithmic idea: grouping rounds of sampling into 'chunks' of n, each arm will be sampled once per chunk, i.e. we perform vector-a-time-sampling. After each chunk of sampling we will check if there is an arm whose anytime lower-confidence bound is larger than the anytime upper-confidence bound of all other arms. On the event that this separation occurs we will stop sampling and return the arm that has become completely separated. It is critical that we make use of anytime confidence bounds here, i.e. confidence bounds that control the probability that deviations of the sample mean *ever* exceed some barrier, rather than just after a fixed number of samples. Omitting this correction would result in a drastic underestimation of the size of the confidence intervals and an algorithm with mis-calibrated guarantees on the quality of its recommendation.

We now turn to the construction of anytime confidence intervals (or confidence sequences) for the sample means of arms that will hold at any round. A reasonably clever way of achieving such a bound is to inflate a Hoeffding confidence interval so that a union bound over an indefinite number of rounds provides the requisite error control:

Lemma 3.6 (Anytime Hoeffding confidence sequence)**.** *Let $S_t = \sum_{s=1}^{t} X_s$ be the sum of t independent 1-subgaussian samples.*

$$\mathbb{P}\left(\exists t \geq 1 : |S_t| \geq \sqrt{2t \log(t(t+1)/\delta)}\right) \leq \delta.$$

Proof. By a union bound over time we have

$$\begin{aligned}
\mathbb{P}\left(\exists t \geq 1 : |S_t| \geq \sqrt{2t \log(t(t+1)/\delta)}\right) &\leq \sum_{t=1}^{\infty} \mathbb{P}\left(|S_t| \geq \sqrt{2t \log(t(t+1)/\delta)}\right) \\
&\leq \delta \cdot \sum_{t=1}^{\infty} \frac{1}{t(t+1)} && \text{(Hoeffding's inequality)} \\
&= \delta \cdot \sum_{t=1}^{\infty} \frac{1}{t} - \frac{1}{t+1} \\
& && \text{(partial fraction decomposition)} \\
&= \delta. && \text{(telescoping the summation)}
\end{aligned}$$

□

[3]As mentioned in Chapter 1 early research into these problems relied on the existence of a lower bound on the smallest gap between arms which effectively reduces to a problem of (ε, δ)-PAC arm identification.

Implementing this form of confidence interval into our proposed algorithm would mean that compared with a classical Hoeffding confidence bound, for a fixed value of δ our confidence intervals would narrow at a slower rate of $\sqrt{\log(t)/t}$ rather than $\sqrt{1/t}$, necessitating more rounds of sampling before an arm's interval dominates and the stopping condition is met. An immediate question is whether or not this confidence sequence is as tight as we can make it.

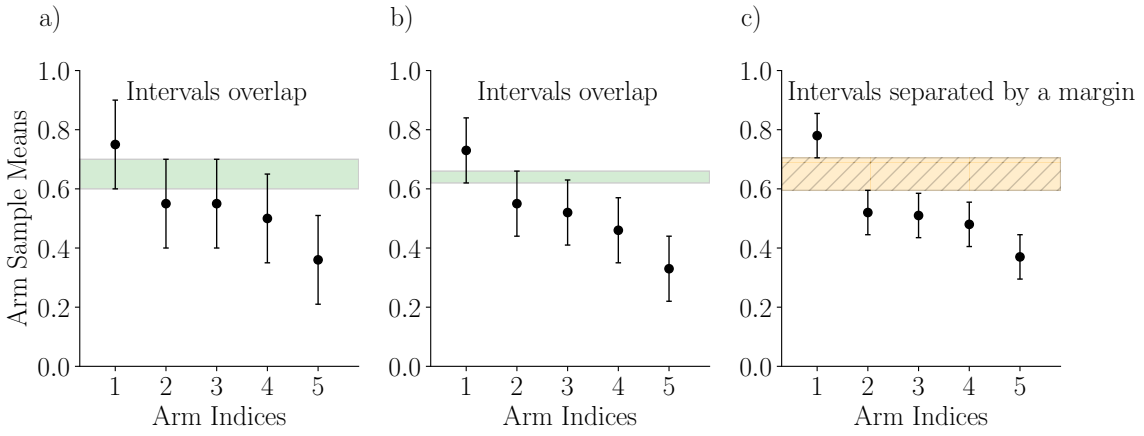

Figure 3.1: An illustration of the behaviour of Uniform Exploration with Adaptive Stopping (Algorithm 2). In a) we see the sample means and confidence intervals after t 'chunks' of uniform sampling ($n \cdot t$ total samples), noting that they are of equal widths due to our vector-at-a-time sampling procedure. Here the intervals for arms 1 and 2 overlap and so we take another sample from each arm. In b), after $t + 1$ rounds of uniform sampling the sample means and confidence intervals have shifted, but a small overlap remains. In c), after $t + 2$ rounds of uniform sampling arm 1 has completely separated from the rest of the arms. At this point Algorithm 2 would stop and return arm 1.

One of the fruits of pure exploration research has been the development of (asymptotically) tight anytime confidence intervals based on a finite time variant of the Law of the Iterated Logarithm presented by Jamieson et al. (2014). We make use of the form that was proposed independently by Kaufmann et al. (2016):

Theorem 3.7 (Kaufmann et al. (2016, Theorem 8))**.** *For* $\delta \leq 0.1$*,*

$$\mathbb{P}\left(\exists t \geq 1 : S_t > \sqrt{2t\beta(t,\delta)}\right) \leq \delta,$$

where

$$\beta(t,\delta) = \log(1/\delta) + 3\log\log(1/\delta) + (3/2)\log\log(et/2).$$

Comparing $\beta(t,\delta)$ with the interval width from Lemma 3.6 we see that for a fixed value of δ we improve to a $\sqrt{\log\log(t)/t}$ width for our confidence intervals, coming within a doubly-

logarithmic factor of the fixed-time result, and provides us with the best performance we can hope for (at least for the two-arm case, as discussed by Jamieson et al. (2014)).

Theorem 3.7 provides us a with a confidence sequence for our algorithm, and making the modification $\delta \to \delta/n$ and taking a union bound over arms gives a bound for uniform exploration.

Define T_t to be the total number of samples taken from all arms at the end of round t.
Require: Confidence level $\delta \in (0, 1)$
1: Set $t = 1$
2: Take one sample from each arm ($T_1 = n$)
3: **while** $\max_{i\in[n]} \hat{\mu}_i(T_t) - 2\sqrt{2\beta(t, \frac{\delta}{2n})/t} \leq \min_{j\in[n]} \hat{\mu}_j(T_t)$ **do**
4: Sample each arm in $[n]$ once
5: $t \leftarrow t + 1$
6: **end while**
7: Return $\psi_{T_t} = \arg\max_{i\in[n]} \hat{\mu}_i(T_t)$, breaking ties in a consistent manner

Algorithm 2: Uniform Exploration with Adaptive Stopping

Theorem 3.8 (Uniform Exploration with Adaptive Stopping)**.** *Run Algorithm 2 on an n-armed 1-subgaussian instance ν. With probability at least $1 - \delta$ the algorithm stops and returns the optimal arm after no more than $O\left(n\Delta_2^{-2}\log(n/\delta) + n\Delta_2^{-2}\log\log(\Delta_2^{-2})\right)$ samples, and $\mathbb{P}_{\nu,\pi}(\tau < infty) = 1$.*

Proof. We give a sketch of the proof here. On the event that all confidence sequences hold (which occurs with probability at least $1 - \delta$ by definition of $\beta(t, \delta)$ and a union bound over all arms), all empirical means are within $\sqrt{2\beta\left(t, \frac{\delta}{2n}\right)/t}$ of their true means on any given round t. It follows that for t large enough the sample means of all arms are within $\Delta_2/2$ of their true mean and the optimal arm is separated from the next best arm(s) by at least Δ_2 by the definition of Δ_2. It remains to analyze how large t must be for this to be true. The main idea is captured by considering the following problem:

Find a value τ so that $\forall t \geq \tau : \log(1/\delta) + \log\log(t) \leq 5t \cdot \Delta_2^2$. Examining the choice

$\tau = \Delta_2^{-2} \log(1/\delta) + \Delta_2^{-2} \log\log\left(\Delta_2^{-2}\right)$ we see that for for all $t \geq \tau$:

$$\begin{aligned}
\log(1/\delta) + \log\log(t) &= \log(1/\delta) + \log\log\left[\Delta_2^{-2}\log(1/\delta) + \Delta_2^{-2}\log\log(\Delta_2^{-2})\right] \\
&\leq \log(1/\delta) + \log\log\left[2\Delta_2^{-2}\log(1/\delta)\right] + \log\log\left[2\Delta_2^{-2}\log\log(\Delta_2^{-2})\right] \\
&\qquad\qquad \text{(log sum inequality)} \\
&\leq 2\log(1/\delta) + 4\log\log(\Delta_2^{-2}) + (\log)^4(\Delta_2^{-2}) \\
&\quad \text{(where } (\log)^3(x) \text{ denotes the iterated logarithm } \log\log\log(x)\text{, etc.)} \\
&\leq 5\left(\log(1/\delta) + \log\log(\Delta_2^{-2})\right) \\
&= 5\tau \cdot \Delta_2^2 \\
&\leq 5t \cdot \Delta_2^2.
\end{aligned}$$

It follows that $\tau = O\left(\Delta_2^{-2}\log(n/\delta) + \Delta_2^{-2}\log\log(\Delta_2^{-2})\right)$ rounds of sampling (respectively $n \cdot \tau$ samples) is sufficient for our first result.

That τ is finite almost-surely follows by observing that by, way of contradiction, on the event that τ is not finite, all sample means must converge to the same point with measure-zero. $\square$

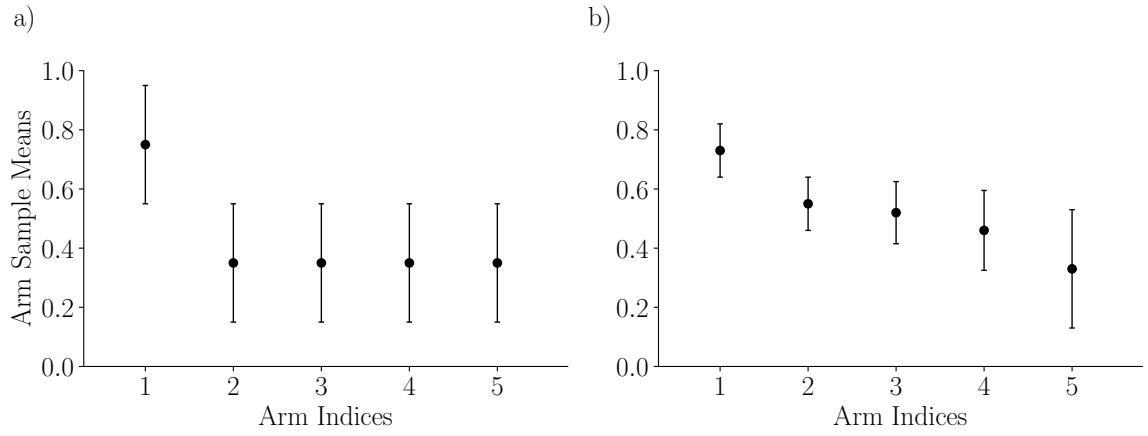

Figure 3.2: When each arm i's confidence interval is no wider than Δ_i (taking $\Delta_1 \coloneqq \Delta_2$) then the unique best arm's confidence interval is disjoint from all other intervals. On an instance with uniform gaps such as in a) this requires uniformly many samples across all arms, whereas for a general instance as in b) we see a larger variation in confidence interval width (and thus numbers of samples required to produce said intervals).

Comparing this result with Theorem 3.1 we find a similar sample complexity, however thanks to the use of a data-dependent stopping time we are able to replace a prescribed $\varepsilon > 0$

with an adaptive Δ_2.[4] On the other hand we also observe that we once again have a factor of $\log(n/\delta)$ from a union bound over arms, which may be prohibitive if n is very large, and as well our result only depends on the smallest gap and so provides worst-case error control over instances where Δ_2 is fixed. Our result comes from requiring that all arms have confidence intervals narrower than Δ_2 so that we can confidently separate the top two arms. If some arms have larger gaps $\Delta_i > \Delta_2$ then on the event that the confidence sequences hold then the intervals of these arms separate from the best arm after only $O\left(\Delta_i^{-2}\log(n/\delta) + \Delta_i^{-2}\log\log(\Delta_i^{-2})\right)$ rounds of sampling for arm i. If we could simply eliminate each arm after it separates from the optimal arm then we would achieve a sample complexity like

$$O\left(\sum_{i=1}^{n}\Delta_i^{-2}\log(n/\delta) + \Delta_i^{-2}\log\log(\Delta_i^{-2})\right),$$

where we take $\Delta_1 \coloneqq \Delta_2$. Fig. 3.2 illustrates this idea. Unfortunately this insight does not provide us with an algorithm since it requires us to know the identity of the optimal arm, but it does provide us with a sample complexity to aim for: $\mathcal{H}\log(1/\delta)$ where $\mathcal{H} \coloneqq \sum_{i=2}^{n}\Delta_i^{-2}$ is an instance dependent complexity term whose use is justified by its appearance in the lower bounds that appear later.

3.2.1 lil' UCB

Rather than creating an explicit elimination algorithm we can combine the tight confidence sequences we have just introduced with an upper confidence bound (UCB) style algorithm that in each round pulls the arm with the largest upper confidence bound. If we assume that the confidence sequences for all arms hold indefinitely then we expect that when a sub-optimal arm i has been pulled sufficiently often that its upper confidence sequence is less than Δ_i then we will stop selecting this arm over the optimal arm (provided a minimum number of pulls on the optimal arm as well). This should hold for all suboptimal arms, leading us to expect that eventually we will only pull the optimal arm. This provides us with both a sampling policy (a greedy policy with respect to upper confidence bounds) along with a new form of stopping

[4] In fact if we modify our algorithm with an additional condition that it stops if it ever reaches round $O\left(\varepsilon^{-2}\log(n/\delta) + \varepsilon^{-2}\log\log(\varepsilon^{-2})\right)$ then we will have a (ε,δ)-PAC algorithm with a sample complexity that scales like $O\left(\frac{n}{(\Delta_2^{\varepsilon})^2}\log\left(\frac{n}{\delta}\right) + \frac{n}{(\Delta_2^{\varepsilon})^2}\log\log((\Delta_2^{\varepsilon})^{-2})\right)$, which is the observation that leads to the nearly instance-optimal PAC algorithm by Karnin et al. (2013).

condition – stop sampling when one arm starts being sampled disproportionately, i.e. using

$$\tau = \inf_{t \in \mathbb{N}} \left\{ t : N_i(t) > a \sum_{j \neq i} N_j(t) \right\} \tag{3.1}$$

for some parameter $a > 0$ to be chosen later, and a selection function $\psi_\tau = \arg\max_{i \in [n]} N_i(\tau)$. This design was introduced by Jamieson et al. (2014) with their lil'UCB algorithm. The authors make use of a slightly different confidence sequence than Theorem 3.7 and their bound features a number of parameters so for clarity we will state their algorithm and bounds using their recommended parameters.

Let $U(t, \delta) \coloneqq 2.2\sqrt{\frac{2.02}{t} \log\left(\log\left(1.01t + 2\right)/(1005^{-1}\delta)^{1/(1.01)}\right)}$.
Require: Confidence level $\delta \in (0, 1)$
1: Sample each arm once, meaning $N_i(t) = 1$ for all $i \in [n]$ and set $t \leftarrow n$
2: **while** $N_i(t) < 1 + 9 \sum_{j \neq i} N_i(t),\ \forall i \in [n]$ **do**
3: $\quad A_{t+1} = \arg\max_{i \in [n]}\ \hat{\mu}_i(t) + U(N_i(t), \delta)$
4: $\quad$ Observe $X_{t+1} \sim \nu_{A_{t+1}}$
5: $\quad t \leftarrow t + 1$
6: **end while**
7: Select $\psi_t = \arg\max_{i \in [n]} N_i(t)$

Algorithm 3: lil'UCB ($\varepsilon = 0.01, \beta = 1, a = 9$) (Jamieson et al., 2014)

Theorem 3.9 (Jamieson et al. (2014, Theorem 2))**.** *With probability at least* $1 - \delta$*,* lil'UCB *stops and returns the optimal arm after at most*

$$O\left(\sum_{i=2}^{n} \Delta_i^{-2} \log(1/\delta) + \sum_{i=2}^{n} \Delta_i^{-2} \log\left(\log\left(c/\Delta_i^2\right)\right)\right)$$

samples, where $c > 0$ *is a universal constant.*

The proof of Theorem 3.9 is intricate, but at a high level involves analyzing the event that the empirical means of arms stay within their anytime confidence sequences, along with a careful analysis of the implication of this event on the stopping time condition that enables an upper bound that – in addition to achieving the desired dependence on the gap structure of the instance – also avoids the $\log(n/\delta)$ dependence we saw in our previous bound. Seeing that a UCB-style algorithm is nearly optimal for best arm identification may come as a surprise, after all in Chapter 2 we mentioned that a trade-off exists between pure exploration and cumulative

regret minimization. The resolution to this apparently confusing situation is that the form of the exploration bonuses in lil'UCB mean that it will experience polynomial cumulative regret (Zhong et al., 2022). On the other hand, Zhong et al. show that a simple parameterized modification of lil'UCB can interpolate between nearly optimal performance for the cumulative regret and BAI objectives.

3.2.2 The Sample Complexity of Best Arm Identification

From our discussion of sample complexity results in the PAC setting we see that a specialization of Theorem 3.5 to $\varepsilon = 0$ tells us that fixing $0 \leq \delta \leq 0.15$, for any algorithm $\mathsf{Alg} = (\pi, \tau, \psi)$ that satisfies $\mathbb{P}_{\nu,\pi}(\psi_\tau \neq 1) \leq \delta$ for an arbitrary 1-subgaussian instance ν we have $\mathbb{E}_{\nu,\pi}[\tau] \geq \sum_{i=1}^{n} \Delta_i^{-2} \log\left(\frac{1}{2.4\delta}\right)$, where we take $\Delta_1 := \Delta_2$ by convention. This result has been known since Mannor and Tsitsiklis (2004). One of the contributions of Jamieson et al. (2014) was their proof that the doubly-logarithmic terms in Δ_i^{-2} as seen in the bound for lil'UCB are necessary, at least in the case of $n = 2$. This left open the possibility that the sample complexity for BAI scales like $\Theta\left(\sum_{i=2}^{n} \Delta_i^{-2} \log(1/\delta) + \Delta_2^{-2} \log\log \Delta_2^{-1}\right)$, where the second term depends only on the smallest gap. This was shown to be optimistic inL. Chen et al. (2017), where a lower bound term of order $\Omega\left(\sum_{i=2}^{n} \Delta_i^{-2} \log\log n\right)$ was revealed. The correct doubly-logarithmic instance dependent complexity has not been resolved at the time of this work.

Track and Stop: Optimal Asymptotic Sample Complexity

We now turn briefly to asymptotic results, specifically the scaling of the instance dependent sample complexity in the limit as $\delta \to 0$. The non-asymptotic lower bounds from earlier show that as $\delta \to 0$ the sample complexity of BAI necessarily increases without bounds on some instances. That is to say that making an infinitely confident decision from stochastic samples with finite dispersion requires an infinite number of samples. This behaviour is expected and not the focus here, rather we are interested in the problem specific rate that the sample complexity grows at as δ goes to zero:

$$\frac{\mathbb{E}_{\nu,\pi_\delta}[\tau_\delta]}{\log(1/\delta)},$$

where here we write $\mathsf{Alg}_\delta = (\pi_\delta, \tau_\delta, \psi_\delta)$ to emphasize that a general algorithm in the fixed confidence setting is instantiated with a fixed value of δ, on which its sampling policy, stopping time and selection function may all depend. We will drop this explicit notation moving

forwards. From Theorem 3.9 we know that we can achieve asymptotic performance that scales like $\mathcal{H} = \sum_{i=2}^{n} \Delta_i^{-2}$ up to a constant factor, however the goal of asymptotically optimal algorithm design is to achieve the optimal scaling *including* an optimal constant factor.

In order to better understand the asymptotic sample complexity we first state a lower bound that is itself a modified version of Garivier and Kaufmann (2016, Theorem 1). Recall that unless stated otherwise we have been considering instances $\nu \in \mathcal{E}_{1-\text{SG}}^n$, the class of n-armed bandit instances with 1-subgaussian arms.

Theorem 3.10 (Lattimore and Szepesvári (2020, Theorem 33.5))**.** *Suppose that for fixed $\delta \in (0,1)$ $\mathsf{Alg} = (\pi, \tau, \psi)$ is δ-correct on $\mathcal{E}_{1-\text{SG}}^n$. Let $i^*(\nu) \coloneqq \arg\max_{i \in \nu} \mu_i(\nu)$ and $\mathcal{E}_{\text{alt}}^n(\nu) = \left\{\tilde{\nu} \in \mathcal{E}_{1-\text{SG}}^n : i^*(\nu) \neq i^*(\tilde{\nu})\right\}$, the set of n-armed 1-subgaussian instances with distinct optimal arms from $i^*(\nu)$. Then we have $\mathbb{E}_{\nu,\pi}[\tau] \geq c^*(\nu) \log(4/\delta)$ where*

$$c^*(\nu)^{-1} \coloneqq \sup_{\alpha \in \mathcal{M}_n} \inf_{\tilde{\nu} \in \mathcal{E}_{\text{alt}}^n(\nu)} \sum_{i=1}^{n} \alpha_i \text{KL}(\nu_i, \tilde{\nu}_i),$$

and where $\mathcal{M}_n$ denotes the probability simplex in dimension n and with the convention that $c^*(\nu)^{-1} = \infty$ when $c^*(\nu) = 0$.

Proof. Let $\mathsf{Alg} = (\pi, \tau, \psi)$ be a δ-correct algorithm for $\mathcal{E}_{1-\text{SG}}^n$ and let $\tilde{\nu} \in \mathcal{E}_{\text{alt}}^n(\nu)$. Then by the definition of a δ-correct learner on $\mathcal{E}_{1-\text{SG}}^n$ we have

$$\begin{aligned}
2\delta &\geq \mathbb{P}_{\nu,\pi}(\psi_\tau \neq i^*(\nu)) + \mathbb{P}_{\tilde{\nu},\pi}(\psi_\tau \neq i^*(\tilde{\nu})) \\
&\geq \mathbb{P}_{\nu,\pi}(\psi_\tau \neq i^*(\nu)) + \mathbb{P}_{\tilde{\nu},\pi}(\psi_\tau = i^*(\nu)) && \text{(since } i^*(\nu) \neq i^*(\tilde{\nu})\text{)} \\
&\geq \frac{1}{2} \exp\left(-\text{KL}\left(\mathbb{P}_{\nu,\pi}, \mathbb{P}_{\tilde{\nu},\pi}\right)\right) && \text{(Bretagnolle-Huber inequality)} \\
&\geq \frac{1}{2} \exp\left(-\sum_{i=1}^{n} \mathbb{E}_{\nu,\pi}[N_i(\tau)] \, \text{KL}(\nu_i, \tilde{\nu}_i)\right). \\
& && \text{(by the divergence decomposition lemma for random stopping times)}
\end{aligned}$$

Tightening the RHS with a judicious choice of $\tilde{\nu}$ we can use this result to show that

$$\begin{aligned}
\log(4/\delta) &\leq \mathbb{E}_{\nu,\pi}[\tau] \cdot \inf_{\tilde{\nu} \in \mathcal{E}_{\text{alt}}^n(\nu)} \sum_{i=1}^{n} \left(\frac{\mathbb{E}_{\nu,\pi}[N_i(\tau)]}{\mathbb{E}_{\nu,\pi}[\tau]}\right) \text{KL}(\nu_i, \tilde{\nu}_i) \\
&\leq \mathbb{E}_{\nu,\pi}[\tau] \cdot \sup_{\alpha \in \mathcal{M}_n} \inf_{\tilde{\nu} \in \mathcal{E}_{\text{alt}}^n(\nu)} \sum_{i=1}^{n} \alpha_i \text{KL}(\nu_i, \tilde{\nu}_i).
\end{aligned}$$

□

This proof reveals that in order for an algorithm to approach the lower bound as $\delta \to 0$ we need $\mathbb{E}_{\nu,\pi}\left[N_i(\tau)\right]/\mathbb{E}_{\nu,\pi}\left[\tau\right]$ to approach $\alpha_i^*(\nu)$ where $\alpha^*(\nu)$ satisfies

$$\inf_{\tilde{\nu}\in\mathcal{E}^n_{\text{alt}}(\nu)} \sum_{i=1}^n \alpha_i^*(\nu)\mathrm{KL}\left(\nu_i, \tilde{\nu}_i\right) = \sup_{\alpha\in\mathcal{M}_n} \inf_{\tilde{\nu}\in\mathcal{E}^n_{\text{alt}}(\nu)} \sum_{i=1}^n \alpha_i \mathrm{KL}\left(\nu_i, \tilde{\nu}_i\right).$$

That is to say that on any given instance ν there exists some optimal proportion of pulls $\alpha^*(\nu)$ which depends on ν that any asymptotically optimal algorithm *must* converge to. Note that this optimal proportion is the one that maximizes the ability of an algorithm to distinguish between ν, and an adversarial alternative instance $\tilde{\nu} \in \mathcal{E}^n_{\text{alt}}(\nu)$.

Let $\hat{\nu}(t)$ to be the bandit instance $\hat{\nu} \in \mathcal{E}$ with mean vector $\hat{\mu}(t)$, $\hat{\alpha}^*(t) = \alpha^*(\hat{\nu}(t))$, $Z_t = \inf_{\tilde{\nu}\in\mathcal{E}^n_{\text{alt}}(\nu)} \sum_{i=1}^n N_i(t)\cdot \mathrm{KL}\left(\hat{\nu}_i(t), \nu_i\right)$ and $\beta_t(\delta) = n\log(t^2+t) + f^{-1}(\delta)$ for $f(x) = \exp(n-x)(x/n)^n$.

Require: Confidence level $\delta \in (0,1)$, bandit class $\mathcal{E}$ such that $\nu \in \mathcal{E}$

Sample each arm once, meaning $N_i(t) = 1$ for all $i \in [n]$ and set $t \leftarrow n$
while $Z_t < \beta_t(\delta)$ **do**
  if $\arg\min_{i\in[n]} N_i(t) \leq \sqrt{t}$ **then**
    $A_{t+1} = \arg\min_{i\in[n]} N_i(t)$ ▷ Forced Exploration
  else
    $A_{t+1} = \arg\max_{i\in[n]} \left(t\hat{\alpha}_i^*(t) - N_i(t)\right)$ ▷ Converge toward $\hat{\alpha}^*(t)$
  end if
  Observe $X_{t+1} \sim \nu_{A_{t+1}}$
  $t \leftarrow t+1$
end while
Select $\psi_t = \arg\max_{i\in[n]} \hat{\nu}(t)$

Algorithm 4: Track-and-Stop (Garivier and Kaufmann, 2016)

The main contribution of Garivier and Kaufmann (2016) is their proof that for the class of single-parameter exponential family bandits Theorem 3.10 is tight, proposing the Track-and-Stop algorithm which matches the asymptotic lower bound exactly.

Theorem 3.11 (Lattimore and Szepesvári (2020, Theorem 33.6))**.** Track-and-Stop *(Algorithm 4) is δ-correct on $\mathcal{E}^n_{\mathcal{N}_1}$, the class of bandit instances with shifted standard normal, and for any $\nu \in \mathcal{E}^n_{\mathcal{N}_1}$ with a unique optimal arm*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\nu,\pi}\left[\tau\right]}{\log(1/\delta)} = c^*(\nu; \mathcal{E}^n_{\mathcal{N}_1}),$$

where (π, τ, ψ) correspond to Algorithm 4 and $c^*(\nu; \mathcal{E}^n_{\mathcal{N}_1})$ corresponds to the result of Theorem 3.10 when we replace $\mathcal{E}^n_{1-SG}$ with $\mathcal{E}^n_{\mathcal{N}}(1)$

Specializing to the shifted standard normal case it is possible to verify that

$$\sum_{i=1}^{n} \frac{2}{\Delta_i^2} \leq c^*(\nu; \mathcal{E}^n_{\mathcal{N}_1}) \leq 2 \sum_{i=1}^{n} \frac{2}{\Delta_i^2}, \tag{3.2}$$

where again we take $\Delta_1 \coloneqq \Delta_2$.

Track-and-Stop is a rare example of a bandit algorithm that uses plug-in estimates and still achieves a performance guarantee. The algorithm follows iterations of estimating the instance that is being played, calculating a stopping time and optimal pull proportions for that estimated instance, and either pulling arms according to the estimates optimal proportions or performing forced exploration of under-sampled arms. This forced-exploration step is crucial as it is well known that the näive use of plug-in estimators can be unstable, potentially never recovering from an initial bad estimate. For instance consider a sampling policy that is greedy with respect to observed sample means. If in early rounds the optimal arm's sample mean drops sufficiently far below its true value then if the sample means of other arms are well behaved then its possible that the algorithm will never pull this arm again, never correcting for this earlier bad estimate. UCB addresses this issue with carefully designed exploration bonuses, and Track-and-Stop uses forced exploration to ensure that the algorithm can recover from a bad estimate for $\alpha^*(\nu)(t)$ and eventually converge to $\hat{\alpha}^*(t) = \alpha^*$. Unlike the algorithms we have seen previously there are no finite-time guarantees for Track-and-Stop. The sampling policy π is independent of the desired confidence level δ, meaning that Track-and-Stop is an inherently asymptotic algorithm – as opposed to an algorithm where the prescribed δ influences the degree of exploration Track-and-Stop targets the limit of infinite samples where the optimal arm allocation ratios α^* are independent of δ. Degenne, Koolen, and Ménard (2019) addressed this issue, explicitly approaching pure exploration as an unknown game and developing an algorithm with finite-time sample complexity guarantees that retains the optimal asymptotic sample complexity on bandits with single-parameter exponential family arms.

3.3 Making Sense of Many Good Arms: Alternative PAC Frameworks

So far in this chapter we have considered problem settings that capture the challenge of confirming that a selected arm matches a PAC criteria with little to no foreknowledge of the instance class apart from subgaussian tails on the rewards of arms. While these settings are of

theoretical and specific practical importance they do not always provide the most useful model for understanding the performance of algorithms in certain scenarios. Consider a 'two-level' instance with m arms with mean $(1-\varepsilon)$, and $n-m$ mean zero arms. If we were to run a reasonable pure exploration algorithm on this instance then we would expect that for larger m, fewer samples are required before the empirically best arm is one of the m ε-good arms with high probability. Unfortunately the (ε, δ)-PAC formulation does not capture this intuition; the lower bound by Kaufmann et al. (2016) indicates that on this instance the expected sample complexity is at least $\Omega(m/\varepsilon^2 \cdot \log(1/\delta))$, growing linearly with m. As pointed out by Katz-Samuels and Jamieson (2020) the issue here is in the notion of 'verification' which is baked into the standard PAC problem formulation. This is perhaps best explained in an example: going back to the two-level instance, if we were to simply select an arm uniformly at random, then the odds of finding an ε-good arm are m/n. With that said, if we want the algorithm to be able to provide a guarantee that the selected arm is in fact ε-good in a general setting, then existing PAC lower bounds (e.g. Theorem 3.5) show that the algorithm cannot confidently make a selection without first confirming that every other arm has a mean which is no more than ε better than the selection. This requires $\Omega([m/\varepsilon^2 + (n-m)] \cdot \log(1/\delta))$ samples.

We will consider two ways to achieve bounds that align with our expectations for these problems. One way will be to drop the verification requirement entirely and instead examine how long it takes for algorithms to begin to return good arms. We will first explore a second idea, which is to consider algorithms equipped with additional prior knowledge about the instance they are interacting with. This idea here is to consider situations where algorithms are able to take advantage of prior knowledge on the class of instances to verify and return a solution faster than is possible in a more general setting. Along these lines Chaudhuri and Kalyanakrishnan (2017) proposed a modified PAC framework to study the design and performance limitations of algorithms when the number of ε-good arms is known in advance.

Definition 3.2 ($(n, m, \varepsilon, \delta)$-PAC Algorithm (Chaudhuri and Kalyanakrishnan, 2017))**.** An algorithm $\mathsf{Alg} = (\pi, \tau, \psi)$ is said to be $(n, m, \varepsilon, \delta)$-PAC if for fixed values of $m < n$ and $\varepsilon > 0$ and any n-armed 1-subgaussian bandit instance ν, τ is finite with probability 1 and $\mathbb{P}_{\nu,\pi}\left(\mu_{\psi_\tau} > \mu_m - \varepsilon\right) \leq \delta$.

With Definition 3.2 we can design a fairly simple algorithm with a sample complexity that matches the intuition we developed earlier.

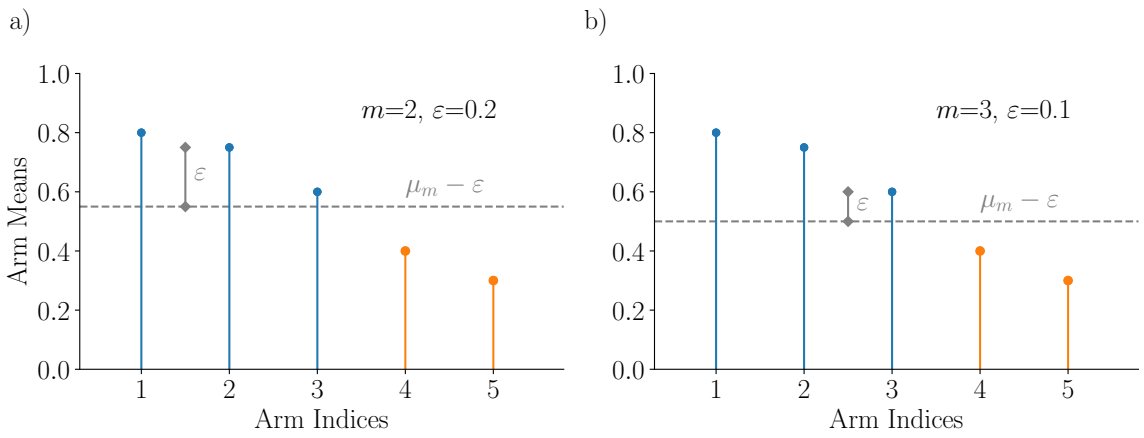

Figure 3.3: An illustrative example of sets of (m, ε)-good arms for different values of m, ε on the same instance. Arms with blue stems are (m, ε)-good while arms with orange stems are not. A pair m, ε defines a threshold on arm means, illustrated here by a dashed grey line.

Require: Confidence level δ, error tolerance $\varepsilon > 0$, problem parameter $1 \leq m < n$

1: Draw a subset of arms $\mathcal{S}$ by drawing $\frac{n}{m} \cdot \log(2/\delta)$ arms from $[n]$ with replacement – if arm i appears more than once in $\mathcal{S}$ then each copy is treated as a new separate arm

2: Select ψ, the result of running Median Elimination (Algorithm 1) on $\mathcal{S}$

Algorithm 5: Subsampling Median Elimination (Chaudhuri and Kalyanakrishnan, 2017)

Theorem 3.12 (Chaudhuri and Kalyanakrishnan (2017, Corollary 3.2))**.** *Algorithm 5 is* $(n, m, \varepsilon, \delta)$*-PAC and satisfies*

$$\mathbb{E}_{\nu,\pi}[\tau] = O\left(\frac{n}{m}\varepsilon^{-2}\log^2(1/\delta)\right).$$

Proof. On a given bandit instance, there are at least m (m, ε)-good arms which satisfy the $(n, m, \varepsilon, \delta)$-PAC condition. In the two-level example from earlier there were exactly m of these arms. This means that if an algorithm selects $n/m \cdot \log(2/\delta)$ arms uniformly at random with replacement, then with probability at least $1 - \left(1 - \frac{m}{n}\right)^{\frac{n}{m}\cdot\log(2/\delta)} \geq 1 - \delta/2$, a good arm is contained in the selection. On this event we know from Theorem 3.3 that running Median Elimination on this subset of arms uses $O(\frac{n}{m} \cdot \frac{1}{\varepsilon^2}\log^2(2/\delta))$ samples to return a good arm with probability at least $1 - \delta/2$, and so a union bound on getting a favourable subsample of arms and the success of Median Elimination shows that this leads to a $(n, m, \varepsilon, \delta)$-PAC algorithm with a sample complexity that scales with n/m instead of n. □

Chaudhuri and Kalyanakrishnan (2017) showed that the sample complexity of Algorithm 5

is within log factors in m, n and an extra $\log(1/\delta)$ of the optimal worst-case sample complexity:

Theorem 3.13 (Chaudhuri and Kalyanakrishnan (2017, Theorem 3.3))**.** *Fix* $0 < \varepsilon \leq \frac{1}{\sqrt{32}}$, $0 < \delta \leq \frac{e^{-1}}{4}$, $m \geq 1$ *and* $n \geq 2m$. *For any algorithm* $\mathsf{Alg} = (\pi, \tau, \psi)$ *that is* $(n, m, \varepsilon, \delta)$*-PAC on* n*-armed instances with rewards in* $[0, 1]$ *almost surely, there exists an* n*-armed instance* ν *for which*

$$\mathbb{E}_{\nu,\pi}[\tau] \geq \frac{1}{306} \cdot \frac{1}{\varepsilon^2} \cdot \frac{n}{m} \log\left(\frac{1}{4\delta}\right).$$

While the $(n, m, \varepsilon, \delta)$-PAC formulation allows us to prove results that better match our understanding of the complexity of problems we want to solve in the real world, the formulation has some practical limitations. A major limitation is that in this setting algorithms require m and ε as inputs to the algorithm in order to reach the sample complexities we expect. This is an unavoidable aspect of the 'verification' step we discussed previously. This leads us to an alternative approach to this problem where, rather than simplifying the verification process by adding prior knowledge, we consider the number of samples required before an algorithm can begin to unverifiably recommend good arms – i.e. without stopping after each recommendation. This metric is used to analyze the behaviour that interacting with an instance induces in pure exploration algorithms.

Definition 3.3 ((ε, δ)-unverifiable sample complexity (Katz-Samuels and Jamieson, 2020, Definition 2))**.** Fix $\varepsilon > 0, \delta \in (0, 1)$, an algorithm $\mathsf{Alg} = (\pi, \tau, \psi)$ and an n-armed 1-subgaussian instance ν. Let $\mathcal{T}_{\varepsilon,\delta}$ be a stopping time with respect to the filtration $\mathcal{F} = (\mathcal{F}_t)_{t \geq 1}$ such that

$$\mathbb{P}_{\nu,\mathsf{Alg}}(\forall t \geq \mathcal{T}_{\nu,\varepsilon,\delta} : \Delta_{\psi_t} < \varepsilon) \geq 1 - \delta. \tag{3.3}$$

Such a stopping time is an (ε, δ)*-unverifiable stopping time of* Alg *with respect to* ν. If for any other $\mathcal{F}$-measurable (ε, δ)-unverifiable stopping time $\mathcal{T}'$ it holds that $\mathcal{T}_{\varepsilon,\delta} < \mathcal{T}'$ almost surely, then $\mathbb{E}_{\nu,\mathsf{Alg}}[\mathcal{T}_{\varepsilon,\delta}]$ is the *expected* (ε, δ)*-unverifiable sample complexity of* Alg *with respect to* ν.

The notion of an 'unverifiable sample complexity' was first introduced into the bandit literature by Katz-Samuels and Jamieson (2020), though the authors of the paper pointed out that the idea of analyzing the unverifiable behaviour of algorithms has been present in the active learning community since Balcan et al. (2010).[5]The essential feature of the unverifiable sample

[5]We remark that the name 'unverifiable sample complexity' is perhaps an unfortunate choice – if we were to

Let $N_i^r(t)$ denote the number of times arm i has been pulled in bracket r up to round t and $\hat{\mu}_i^r(t)$ denote the empirical mean of arm i in bracket r from samples up to round t and let $U(t,\delta) = c\sqrt{\frac{1}{t}\log\left(\log(t)/\delta\right)}$ be an anytime confidence bound e.g. as in Algorithm 3 that satisfies $\mathbb{P}\left(\forall t \in \mathbb{N}, |\hat{\mu}_i(t) - \mu_i| \leq U(t,\delta)\right) \geq 1-\delta$ for $i \in [n]$.

Require: Confidence level δ

$\ell = 0, R_0 = 0, S_0 = \emptyset$
for $t = 1, 2, \ldots$ **do**
  if $t \geq 2^\ell \ell$ **then**
    Draw set of arms $\mathcal{A}_{\ell+1} \sim \binom{[n]}{M_{\ell+1}}$, $M_\ell \coloneqq n \wedge 2^\ell$
    $\ell = \ell + 1$
  end if
  $R_t = 1 + R_t \cdot \mathbb{1}\{R_t < \ell\}$
  if $\exists i \in \mathcal{A}_{R_t} \setminus \mathcal{S}_t$ such that $N_i^{R_t}(t) = 0$ **then**
    Pull $A_{t+1} \in \left\{i \in \mathcal{A}_{R_t} \setminus \mathcal{S}_t : N_i^{R_t}(t) = 0\right\}$
  else
    Pull $A_{t+1} = \arg\max_{i \in \mathcal{A}_{R_t} \setminus \mathcal{S}_t} \hat{\mu}_i^{R_t}(t) + U(N_i^{R_t}(t), \delta)$
  end if
  Observe $X_{t+1} \sim \nu_{A_{t+1}}$
  $\psi_t = \arg\max_{i \in \mathcal{S}_r \text{ for some } r \leq \ell} \hat{\mu}_i^{R_t}(t) - U\left(N_i^{R_t}(t), \frac{\delta}{|\mathcal{A}_r| r^2}\right)$
end for

Algorithm 6: Bracketing UCB Algorithm, ε-good arm identification (Katz-Samuels and Jamieson, 2020)

complexity is that while $\mathcal{T}_{\varepsilon,\delta}$ is a stopping time, it is in general *unknowable* to the algorithm (since it depends on the instance ν). This means that $\mathcal{T}_{\varepsilon,\delta}$ is purely an analysis variable and as such lower bounds on $\mathcal{T}_{\varepsilon,\delta}$ cannot be constructed from the usual reduction to hypothesis testing that would give the $m \cdot \varepsilon^{-2}$ bound we are trying to avoid on instances with m ε-good arms. The authors provided a lower bound on this modified notion of sample complexity:

Theorem 3.14 (Katz-Samuels and Jamieson (2020, Theorem 1))**.** *Fix $\varepsilon > 0, \delta \in (0, 1/16)$. Let ν be a shifted standard normal bandit instance with means $\mu \in \mathbb{R}^n$ with $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$ and let $m = |\{i \in [n] : \mu_i > \mu_1 - \varepsilon\}|$, the number of ε-good arms on the instance. Consider an arbitrary algorithm* Alg $= (\pi, \tau, \psi)$*. For every permutation $\sigma \in \Sigma^n$ let $(\mathcal{F}_t^\sigma)_{t\geq 1}$ be the filtration induced by the interaction of* Alg *with instance $\sigma(\nu)$ with arm indices permuted by σ and let $\mathcal{T}_\sigma$*

restrict our attention to a sufficiently narrow classes of instances ν, then these 'unverifiable stopping times' are possibly known to the algorithm. For example, on the familiar set of n-armed, two-level 1-subgaussian instances with m mean ε arms, and $n - m$ mean zero arms, where $\varepsilon > 0$ is fixed and known in advance.

be a stopping time with respect to $(\mathcal{F}_t^\sigma)_{t\geq 1}$ *for which* $\mathbb{P}_{\sigma(\nu),\pi}\left(\forall t \geq \mathcal{T}_\sigma, \Delta_{\psi_t} < \varepsilon\right) \geq 1-\delta$. *Then*

$$\mathbb{E}_{\sigma\sim\Sigma^n}\,\mathbb{E}_{\sigma(\nu),\mathsf{Alg}}\left[\mathcal{T}_\sigma\right] \geq \mathcal{H}_{\text{low}}(\varepsilon) \coloneqq \frac{1}{64}\left(-\Delta_{m+1}^{-2} + \frac{1}{m}\sum_{i=m+1}^{n}\Delta_i^{-2}\right)$$

Considering a two-level instance with m arms with mean ε and $n-m$ mean zero arms we see that this lower bound implies $\mathcal{H}_{\text{low}}(\varepsilon) \geq \frac{1}{64}\left(\frac{n-2m}{m}\cdot\varepsilon^{-2}\right)$, which for $n \gg m$ provides the expected reduction in sample complexity with m. Furthermore Katz-Samuels and Jamieson (2020) introduced the Bracketing UCB (BUCB) algorithm (Algorithm 6) that comes within logarithmic factors of the lower bound on our two-level setting. The general upper bounds presented in the paper are highly technical and so we present the author's result for the special two-level instance class. Define $\overline{\log}(x) \coloneqq \max\left(\log(x), 1\right)$ and once again taking $m = |\{i \in [n] : \mu_i > \mu_1 - \varepsilon\}|$ we define $\bar{\mathcal{H}}(\varepsilon) \coloneqq \frac{n}{m}\varepsilon^{-2}\log(1/\delta)$.

Theorem 3.15 (Katz-Samuels and Jamieson (2020, Theorem 2))**.** . *Let* $\delta \leq 0.025$ *and* $\varepsilon > 0$. *Let* $\mathcal{F} = (\mathcal{F}_t)_{t\geq 1}$ *be the filtration generated by playing Algorithm 6 on a two-level 1-subgaussian instance* ν *with* m *with means* ε *and* $n-m$ *mean zero arms. Then there exists an* (ε,δ)*-unverifiable stopping time* $\mathcal{T}$ *w.r.t.* $\mathcal{F}$ *such that*

$$\mathbb{E}_{\nu,\pi}\left[\mathcal{T}\right] = \tilde{O}\left(\bar{\mathcal{H}}(\varepsilon)\log\left(\bar{\mathcal{H}}(\varepsilon)\right)\right),$$

where $\tilde{O}$ *hides doubly-logarithmic terms as well as constant factors.*

In a sense Bracketing UCB is an adaptive version of Subsampling Median Elimination (Algorithm 5) that replaces the Median Elimination subroutine with lil'UCB (Algorithm 3). The algorithm operates on increasingly larger subsets, or 'brackets' of arms and within each bracket attempts to identify the largest arm using lil'UCB. Of course the algorithm cannot guarantee that a good arm is contained in a given bracket, but the near-optimal performance of the algorithm can be explained by observing that if m is large then there is a high chance that earlier, smaller brackets contain a good arm that the algorithm can begin to select. On the other hand if m is small then any algorithm will necessarily need closer to order $O(n)$ samples to find an ε-good arm, and BUCB hedges against this second case by eventually considering large enough brackets to enclose a good arm in this case, paying a logarithmic cost in order to adapt to the true value of m.

It is important to note that unlike the result in Theorem 3.12, BUCB achieves this sample complexity without requiring m or ε as an input, using the ε,δ-unverifiable sample complexity

to show that such an algorithm is in fact able to adapt to the complexity of a given instance to obtain the expected rates. It is worth keeping in mind that BUCB is a fixed confidence algorithm with a sampling policy that depends on δ. This means that in practice one can increase δ mid-run and still retain theoretical guarantees, but increasing the confidence (that is, $1 - \delta$) the algorithm operates at mid-run by decreasing δ would invalidate the anytime confidence intervals in the original analysis of the algorithm.

3.4 Aside: Instance-Optimal Rates for Stochastic Convex Optimization

Tools developed to advance theory in the pure exploration in the multi-armed bandit setting also have applications in adjacent problems. Online (stochastic) convex optimization (OCO) is a version of a bandit problem with continuous arms where a learner is tasked with finding an optimizer for a convex function f by interacting with a first-order stochastic 'oracle' that, when queried with a point $x \in \mathcal{D}$ from a convex domain $\mathcal{D}$ returns an unbiased estimate of a sub-gradient of f at x. Definition 3.4 outlines the interaction protocal. The goal of the learner is typically to return a point $\psi_\tau \in \mathcal{D}$ that minimizes the expected sub-optimality of the point or some other application specific measure of the quality of the returned point. As we have come to expect from the multi-armed bandit setting, the sample complexity required to return a good solution depends on the instance that the learner is optimizing, typically related to the curvature of the function around its minimum region. The worst-case sample complexity for these problems has been well understood since Nemirovsky and Yudin (1983), however recently developed algorithms have been shown to able to adapt to the specific hardness of the instance they are optimizing (e.g. Hazan et al. (2007) and Moulines and Bach (2011)) and worst-case analysis is unsuitable for studying the optimality of these methods.

Yuancheng Zhu et al. (2016) introduced the first instance dependent measure of complexity for this problem setting using the idea of *local minimax complexity*. This complexity measure lower bounds the performance of an optimal learner that has to optimize f or an adversarially chosen function g that is allowed to depend on f as well as the choice of learner. Compared with the minimax risk in Chapter 2 local minimax risk considers a more powerful learner that knows it faces one of two functions, rather than only knowing the class of the function it must

optimize.[6] This form of instance-dependent complexity measure differs from those we have discussed so far in this thesis. Rather than considering a restricted class of learners such as (ε, δ)-PAC algorithms and proving a lower-bound on a given instance for algorithms in this class the approach is to instead consider the impossibility of a 'super-efficiency' phenomenon: if an algorithm out-performs the local minimax complexity on an instance f, then there is a closely related function $\tilde{f}$ that is closely related to f for which the algorithm performs worse than the local minimax complexity of $\tilde{f}$. The authors justify the use of this complexity measure with an algorithm that is shown to match the local minimax rate under certain conditions. This setup is in many ways an implicit analogue of the instance-dependent lower bounds that we have seen for bandit algorithms, which leads us to consider whether we can obtain similar results in this setting with an approach we have seen earlier in this chapter.

By adapting tools from the lower bound derived by Garivier and Kaufmann (2016) we provide a parallel result to Yuancheng Zhu et al. (2016) with which we can understand the performance of instance-optimal algorithms in the OCO framework in some settings.

Definition 3.4. Online Convex Optimization with Fixed Confidence

Require: Confidence level δ, loss function $\ell : \mathcal{D} \to \mathbb{R}^+$, access to subgradient oracle

while $t \leq \tau$ **do**
    Select query point $X_t \sim \pi_t(\mathcal{F}_{t-1})$
    Observe noisy subgradient estimate $G_t = \partial f(X_t) + \eta_t$
    (Optionally) Output candidate optimizer $\psi_t : \mathcal{F}_t \to \mathbb{R}^d$
end while
Receive loss $\ell(\psi_\tau, \nu)$

From here on we assume that $\eta_t \sim \mathcal{N}(0, I_d)$ is zero mean i.i.d. noise and that we are seeking to optimize a convex function on $\mathbb{R}^d$, $d \geq 1$. Definition 3.4 is an (ε, δ)-PAC formulation of the online convex optimisation problem that is often viewed in a fixed budget setting with the goal minimizing the expected suboptimality of the optimizer returned after T queries to a noisy first order oracle.

Definition 3.5. Let $\mathcal{C}$ be a collection of convex functions with convex domain $\mathcal{D}$. Fix $\varepsilon > 0, \delta \in (0, 1)$. An algorithm $\mathsf{Alg} = (\pi, \tau, \psi)$ is (ε, δ)-PAC on $\mathcal{C}$ if for all $f \in \mathcal{C}$ $\mathbb{P}_{f,\pi}(\Delta_f(\psi_\tau) \geq \varepsilon) \leq \delta$, where $\Delta_f(x) \coloneqq f(x) - \inf_{\tilde{x} \in \mathcal{D}} f(\tilde{x})$.

[6] If $\mathcal{C} = \{f, g\}$ only has two elements then these complexity measures coincide.

Definition 3.5 provides us with the class of algorithms that our lower bound will apply to. we will require two more definitions first: let $\chi_f^*(\varepsilon) \coloneqq \{x \in \mathcal{D} \mid \Delta_f(x) < \varepsilon\}$ be the set of ε-good minimizers of f and define $\mathcal{E}_{\text{alt}}(f, \varepsilon) \coloneqq \left\{g \in \mathcal{H} \mid \chi_g^*(\varepsilon) \bigcap \chi_f^*(\varepsilon) = \emptyset\right\}$, the subclass of $\mathcal{C}$ for which the ε-good regions of f and the ε-good region of any function in the subclass are disjoint. We omit the dependence on $\mathcal{C}$ here but it will be important to keep in mind when using these definitions to evaluate results.

Theorem 3.16 (Lower bound for PAC Online Stochastic Convex Optimization)**.** *Let $\varepsilon > 0, \delta \in (0, 1)$ and fix some function class $\mathcal{C}$ over convex domain $\mathcal{D}$. We let ∂f denote a subgradient of f that is understood to be chosen in a deterministic manner by the oracle such that $\partial f(x)$ is the same across all queries at point x. For any algorithm $\mathsf{Alg} = (\pi, \tau, \psi)$ that is (ε, δ)-PAC on $\mathcal{C}$, for arbitrary $f \in \mathcal{C}$*

$$\mathbb{E}_{\pi, f}[\tau] \geq c_f^*(\varepsilon) \cdot \log\left(\frac{1}{4\delta}\right),$$

where

$$c_f^*(\varepsilon)^{-1} = \inf_{g \in \mathcal{E}_{alt}(f, \varepsilon)} \sup_{x \in \mathcal{D}} \frac{\|\partial f(x) - \partial g(x)\|_2^2}{2}.$$

The quantity $c_f^*(\varepsilon)$ can be understood as a measure for how ambiguous the function f appears to an algorithm that only receives noisy sub-differential feedback. An adversary attempting to confuse the algorithm would replace f with a function g that is sufficiently similar to f in terms of its gradient (so that the algorithm cannot quickly distinguish between the two functions with first-order derivative information), but for which optimizing g necessarily results in an unsatisfactory solution for f. $c_f^*(\varepsilon)$ plays an analogous role to $\mathcal{H}_\varepsilon(\nu) \coloneqq \sum_{i=2}^n (\Delta_i^\varepsilon)^{-2}$ from Section 3.1.2. The proof of Theorem 3.16 can be found in Section B.1. In order to show that a lower bound is meaningful we must be able to demonstrate that it is tight in some cases.

3.4.1 Results for Various Function Classes

Example 3.1 (Quadratics)**.** Consider a quadratic function class in one dimension $\mathcal{C} = \left\{f : f(x) = \frac{a(x-c)^2}{2},\ a > 0, c \in [-1, 1]\right\}$ with domain $\mathcal{D} = [-1, 1]$. For simplicity consider a lower bound for optimizing $f(x) = ax^2/2$. In this case $\mathcal{E}_{\text{alt}}^n(f, \varepsilon)$ consists of functions $g(x) = \tilde{a}(x - \tilde{c})^2/2$ for which $|\tilde{c}| > \sqrt{2\varepsilon/a} + \sqrt{2\varepsilon/\tilde{a}}$, from which we can verify that with the choice $\tilde{a} = a$ we have $\sup_{x \in [-1,1]} |f'(x) - g'(x)| = 2\sqrt{2a \cdot \varepsilon}$ and $c_f^*(\varepsilon)^{-1} = 4a\varepsilon$ leading to a

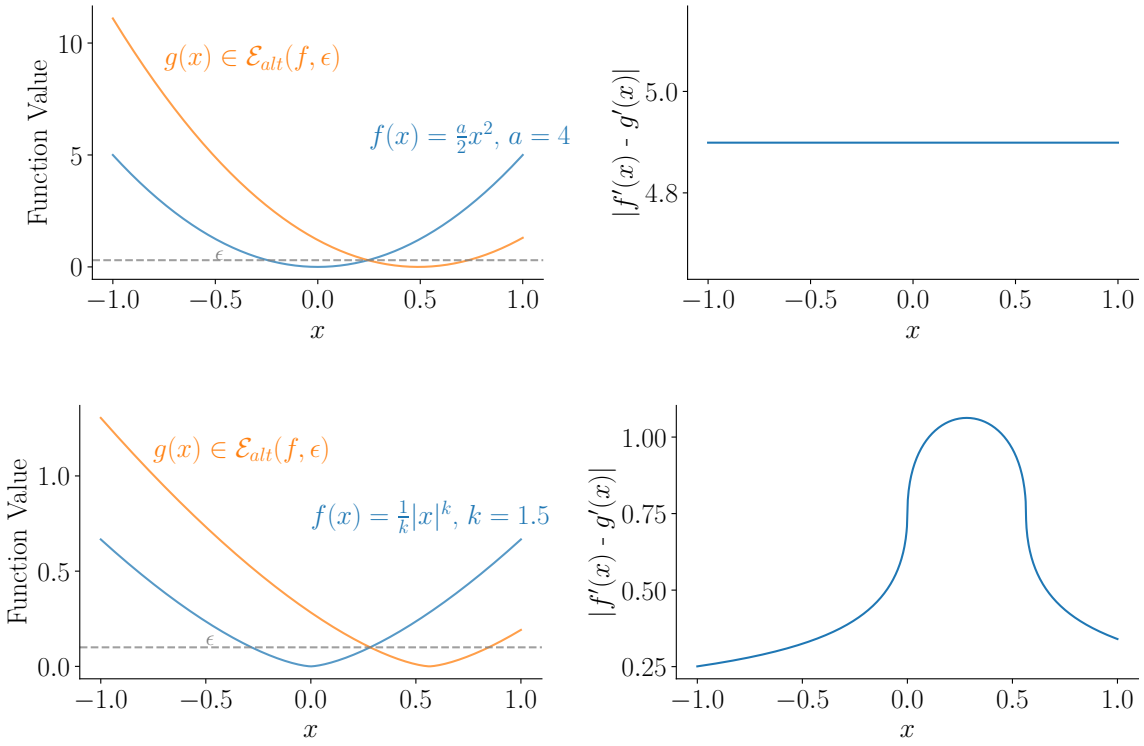

Figure 3.4: Examples of function f and adversarial chosen functions g chosen so that their ε-optimal regions have no overlap but the subgradients of the functions are are close as possible on the domain. The upper figure shows the relatively straightforward case for quadratic functions $f(x) = a(x-c)^2/2$ for some $a > 0, c \in \mathcal{D}$, and the lower figure shows an example for the generalized absolute value function class $f(x) = \frac{1}{k}|x-b|^{1/k}$ for $k \geq 1, b \in \mathcal{D}$.

lower bound of

$$\mathbb{E}_{f,\pi}[\tau] \geq \frac{1}{4a \cdot \varepsilon} \log\left(\frac{1}{4\delta}\right).$$

This corresponds to a result by Yuancheng Zhu et al. (2016) where the authors point out that this rate matches results for algorithms that adapt to strong convexity, e.g. Moulines and Bach (2011).

Example 3.2 (Generalized Absolute Value Class)**.** We can also analyze a function class without uniform strong convexity such as the generalized absolute value class $\mathcal{C} = \left\{f : f(x) = \frac{1}{k}|x-b|^k,\ k \in (1,2], b \in [-1,1]\right\}$ where again we take $\mathcal{D} = [-1,1]$. Let $f(x) = \frac{1}{k}|x|^k$. Then the subclass $\mathcal{E}^n_{\text{alt}}(f,\varepsilon)$ amounts to the set of parameters $\tilde{k}, \tilde{b}$ with $|\tilde{b}| \geq (k\varepsilon)^{1/k} + (\tilde{k}\varepsilon)^{1/\tilde{k}}$. For simplicity in the analysis let $|\tilde{b}| = (k\varepsilon)^{1/k} + (\tilde{k}\varepsilon)^{1/\tilde{k}}$. For $\varepsilon < \frac{1}{8}$ we see that $|\tilde{b}| < 1$, and so we can verify that we have a minimizer for g, x^*_g satisfying $x^*_g \in \chi^*_g(0) \subset [-1,1]$, $\forall \tilde{k} \in (1,2]$. Setting $\tilde{k} = k$ we have

$$\begin{aligned}
\inf_{g \in \mathcal{E}_{\text{alt}}(f,\varepsilon)} \sup_{x \in [-1,1]} |f'(x) - g'(x)| &\leq \sup_{x \in [-1,1]} \left|\partial_x \left(\frac{1}{k}|x|^k - \frac{1}{k}\left|x - 2(k\varepsilon)^{1/k}\right|^k\right)\right| \\
&= \left|\partial_x \left(\frac{1}{k}|x|^k - \frac{1}{k}\left|x - 2(k\varepsilon)^{1/k}\right|^k\right)\right|_{x=(k\varepsilon)^{1/k}} \\
&= 2(k\varepsilon)^{(k-1)/k}.
\end{aligned}$$

leading to the bound

$$\mathbb{E}_{f,\pi}[\tau] \geq \frac{1}{2}(k\varepsilon)^{\frac{-2(k-1)}{k}} \log\left(\frac{1}{4\delta}\right).$$

This result matches the analogous bound by Yuancheng Zhu et al. (2016), who provide the local minimax complexity of minimizing the distance of the returned point to the minimizing region of f, $\text{err}_f(\psi_T) \coloneqq \inf_{x \in \chi^*_f} |x - \psi_T|$, which is related related to the suboptimality of the returned point by $\Delta_{\psi_T} = \frac{1}{k}|\text{err}_f(\psi_T)|^{1/k}$. Furthermore, an algorithm by Ramdas and Singh (2013) matches our lower bound up to log factors with an algorithm with a fixed stopping time that achieves a suboptimality which scales like $\varepsilon = \tilde{O}\left(T^{-\frac{k}{2(k-1)}}\right)$ for fixed δ, compared with our result, which says that

$$\varepsilon = \Omega\left[\left(k^{-1} \cdot 2^{-\frac{k}{2k-2}}\right) \cdot T^{-\frac{k}{2k-2}}\right]. \tag{3.4}$$

Chapter 4

Pure Exploration with a Fixed Budget

Pure exploration with a fixed budget is perhaps the simplest setting to conceptualize. In this setting the goal is to design algorithms that make the best decision possible after a finite sample budget T. This setting lends itself to situations where an experimenter needs to make the most out of the finite resources, be it an advertising budget as in introduction of this thesis, a computational budget when evaluating a collection of supervised machine learning models, or essentially any scenario with a simulator and a hard budget cap.

Definition 4.1. Pure Exploration with a Fixed Budget

Require: Sampling budget $T \in \mathbb{N}$, loss function $\ell : [n] \to \mathbb{R}^+$
1: **while** $t \leq T$ **do**
2: Select arm $A_t \sim \pi_t(\mathcal{F}_{t-1})$
3: Observe reward $X_t \sim \nu_{A_t}$
4: (Optionally) Output $\mathcal{F}_t$-measurable arm selection ψ_t
5: **end while**
6: Receive loss $\ell(\psi_T, \nu)$

Tight analysis for pure exploration algorithms in the fixed budget setting has proved more challenging than for their fixed-confidence counterparts. This is in part due to the need for a precise characterization of the exploration-exploitation trade-off present in this problem. In Chapter 3 we discussed the Track-and-Stop algorithm that achieves the asymptotically optimal sample complexity for the BAI objective. Track-and-Stop is based around ideas derived from a matching lower-bound: how should an algorithm sample if it had full knowledge of the instance that it is interacting with? What loss would this optimal learner incur? The algorithm takes a straightforward approach to achieving the lower bound. Given the samples that the algorithm has seen so far, further samples are drawn according to a balance between producing

a more accurate estimate of the instance, and sampling arms according to the asymptotically optimal proportions for the current estimate of the instance. The presence of a finite, fixed budget complicates this design principle. Whereas PAC algorithms have no choice other than to sufficiently resolve the uncertainty about the gap structure in order to make a correct decision, fixed budget algorithms can be understood to be balancing an exploration-exploitation trade-off between sampling to collect information about the gap structure of the instance, and sampling as if these estimates are the ground truth. Interestingly we will see that if an algorithm is given enough information up front about the instance its interacting with, then results from the fixed confidence setting transfer. In contrast, in the general case this exploration-exploitation trade-off can make the sample complexity of fixed budget pure exploration larger than in the fixed confidence setting, where we understand the 'sample complexity' to mean the minimum size of the budget T required to reduce the probability of an incorrect selection below some probability.

4.1 Best Arm Identification

In the fixed budget best arm identification (BAI) problem we are interested in minimizing $\mathbb{P}_{\nu,\pi}\left(\psi_T \neq 1\right)$ subject to a finite sampling budget T, where once again in this chapter we will work under the assumption that a unique optimal arm exists. Surprisingly this simple setup was not studied in depth prior to Audibert, Sebastien Bubeck, and Remi Munos (2010).

As in the previous chapter let us once again consider a uniform exploration algorithm (UE) in order to provide a performance benchmark for comparison with results from adaptive algorithms. In this setting there is no room for clever sample size selections or stopping rules. We will simply take T/n samples from each arm and return $\psi_T = \arg\max_{i\in[n]} \hat{\mu}_i(T)$.

Lemma 4.1. *Uniformly drawing T/n samples from each arm and selecting the empirical best arm (=(π, ψ)) on instance ν gives*

$$\mathbb{P}_{\nu,\pi}\left(\psi_T \neq 1\right) \leq \sum_{i=2}^{n} \exp\left(-\frac{\Delta_i^2 T}{4n}\right).$$

Proof. An application of the Chernoff-Hoeffding inequality for 1-subgaussian random variables

and a union bound yields

$$\begin{aligned}\mathbb{P}_{\nu,\pi}\left(\psi_T \neq 1\right) &= \sum_{i=2}^{n} \mathbb{P}_{\nu,\mathrm{UE}}\left(\psi_T = i\right) \\ &\leq \sum_{i=2}^{n} \exp\left(-\frac{\Delta_i^2 T}{4n}\right). \qquad \text{(Hoeffding)}\end{aligned}$$

□

Noticing that $\sum_{i=2}^{n} \exp\left(-\Delta_i^2 T/n\right) \leq n \exp\left(-\Delta_2^2 T/n\right)$ we see that if all the arm gaps are uniform then all terms in the summation above are balanced and uniform exploration achieves error control like $\exp\left(-\Delta^2 T/n\right)$, which we can see from Theorem 4.2 turns out to be essentially the optimal error rate for these instances.

Theorem 4.2. *Let ν be a Bernoulli bandit instance with $k \geq 10$ arms, and arm means $\mu_1 = 1/2 + \Delta$,p $\left(\mu_i\right)_{i=2}^{k} = 1/2$ for some $\Delta \leq 1/4$. For any learner $\mathsf{Alg} = (\pi, \psi)$ subject to sampling budget T, there exists some permutation $\sigma \in \Sigma^n$ such that*

$$\mathbb{P}_{\sigma(\nu),\pi}\left(\psi_T \neq \sigma(1)\right) \geq \frac{2}{3(n-1)} \exp\left(-\frac{3\Delta^2 T}{(n-1)} - 2\log 3\sqrt{\frac{T}{(n-1)}\log(6n^2)}\right).$$

It follows that when $T \geq \left(\frac{2\log 3}{3}\right)^2 n \cdot \Delta^{-4} \log(6n^2)$, we have

$$\mathbb{P}_{\sigma(\nu),\pi}\left(\psi_T \neq \sigma(1)\right) \geq \frac{2}{3n-1} \cdot \exp\left(-\frac{6\Delta^2 T}{n-1}\right).$$

The proof of Theorem 4.2 can be found in Section C.1.

We will see later that this effectively corresponds to a worst-case instance for BAI, and demonstrates that just as in the fixed confidence setting, a non-adaptive algorithm is sufficient for nearly-optimal worst-case performance on pure exploration. As in the fixed confidence setting if some of the arm gaps are much smaller than others then more adaptive algorithms have the ability to allocate more samples towards distinguishing small gaps at the expense of sampling from arms that are clearly sub-optimal.

Working towards increasingly adaptive algorithms let us consider an algorithm that is allowed to 'cheat' in the sense that the algorithm has knowledge of the suboptimality gap structure $\{\Delta_i\}_{i=1}^{n}$ in advance, but not the actual correspondence between arm indices and their gaps. In this situation a natural strategy arises: we know that with $O(\Delta_i^{-2})$ samples from each

Let $t_i = \Delta_{n-i+1}^{-2} \cdot T/(\mathcal{H} + \Delta_2^{-2})$ for $i \in [n]$, $t_0 = 0$. Define $\hat{\mu}_i^\ell$ to be the mean of samples from rounds up to and including round ℓ for arm $i \in [n]$.

Require: Sample budget $T \geq \mathcal{H} + \Delta_2^{-2}$, $\mathcal{H} = \sum_{i=2}^n \Delta_i^{-2}$

1: $\mathcal{S}_0 \leftarrow [n]$
2: **for** round $\ell \in [1 : n-1]$ **do**
3: Sample arms in $\mathcal{S}_\ell$ uniformly with $t_\ell - t_{\ell-1}$ samples each
4: $\mathcal{S}_{\ell+1} = \mathcal{S}_\ell \setminus \arg\min_{i \in \mathcal{S}_\ell} \hat{\mu}_i^\ell$
5: **end for**
6: Return the surviving arm $\psi_T \in \mathcal{S}_n$

Algorithm 7: Fixed Budget Successive Elimination with Known Biases (FB-SE)

arm if confidence intervals hold for all arms then arm 1 will have a larger mean than arm $i > 1$, and so if we take $O(\Delta_n^{-2})$ samples from each arm and remove the worst performing arm we are likely to keep the best arm. Since we know the gaps we can simply repeat this process of elimination while ensuring that we use enough samples to keep the sample mean of arm 1 above that of the worst arm in the current phase. Algorithm 7 is inspired by Successive Elimination Known Biases, a strategy introduced by Even-Dar et al. (2002) for the PAC setting.

Theorem 4.3. *Running Algorithm 7 (FB-SE$= (\pi, \psi)$) on instance ν with sampling budget $T \geq \mathcal{H} + \Delta_2^{-2}$ results in*

$$\mathbb{P}_{\nu,\pi}\left(\psi_T \neq 1\right) \leq n^2 \cdot \exp\left(-\frac{T}{8\mathcal{H}}\right).$$

Proof. The proof consists of two parts. First we will prove that the best arm is not eliminated with probability at least $1 - n^2 \exp\left(-\frac{T}{8\mathcal{H}}\right)$. We then show that the algorithm does not exceed the sampling budget T.

Correctness: Consider the reward table probability model where in round t when a learner samples arm $i \in [n]$ they observe $X_{i,t} \sim \nu_i$ from a table X of pre-drawn values. This allows us to define $\hat{\mu}_i^\ell := \frac{1}{N_i^\ell} \sum_{t=1}^{N_i^\ell} X_{i,t}$, where $N_i^\ell := \dfrac{T}{\Delta_{n-\ell}^2 \cdot \left(\mathcal{H} + \Delta_2^{-2}\right)}$. It follows that at the end of phase ℓ the sample mean of arm i considering all samples up to round ℓ is $\hat{\mu}_i^\ell$. We will work on the event

$$\mathcal{E} = \left\{\forall \ell \in [1 : n-1], i \in [n],\ |\hat{\mu}_i^\ell - \mu_i| < \Delta_{n-\ell+1}/2\right\} \tag{4.1}$$

which by Hoeffding bounds and a union bound over all arms and rounds of elimination, occurs with probability at least $1 - n^2 \exp\left(-\frac{T}{8\mathcal{H}}\right)$, where we use $\mathcal{H} + \Delta_2^{-1} \leq 2\mathcal{H}$. All we need to

do is show that $\mathcal{E}$ is sufficient for correctness. At the start of phase $\ell \in [0 : n-2]$ we have eliminated ℓ arms. It follows that in phase ℓ there must exist some arm $j \in [n]$ such that $\Delta_j \geq \Delta_{n-\ell}$. On $\mathcal{E}$ it holds that

$$\hat{\mu}_1^\ell - \hat{\mu}_j^\ell = \left(\hat{\mu}_1^\ell - \mu_1\right) - \left(\hat{\mu}_j^\ell - \mu_j\right) + \Delta_j > \Delta_j - \Delta_{n-\ell} \geq 0$$

where the second last inequality holds on the event $\mathcal{E}$ and the last inequality follows from the definition of $j \in [n]$. It follows that on event $\mathcal{E}$ arm 1 is selected.

Budget Constraint: By our choice of the sampling budget of each round, the arm which is eliminated after phase ℓ (where define the last arm in S_n as being 'eliminated' after it is selected) is pulled t_ℓ times in total. Hence the total number of samples drawn is

$$\sum_{\ell=1}^{n} t_\ell = \frac{T}{\mathcal{H} + \Delta_2^{-2}} \cdot \sum_{\ell=1}^{n} \Delta_{n-\ell+1}^2 = \frac{T}{\mathcal{H} + \Delta_2^{-2}} \cdot \left(\Delta_2 + \sum_{\ell=2}^{n} \Delta_\ell^{-2}\right) = T, \tag{4.2}$$

where we recall that we make use of the convention that $\Delta_1 = \Delta_2$. □

Successive Elimination is an algorithm with an intuitive exploration scheme and a sample complexity of $O\left(\mathcal{H} \log(n/\delta)\right)$ for identifying the best arm, which appears familar coming from Chapter 3, where lower and upper bounds (Theorem 3.5 and Theorem 3.9) bracket a matching sample complexity up to doubly-logarithmic terms in the gaps of the instance. These parallel results diverge when we move to consider algorithms that are sample-efficient when the value of $\mathcal{H}$ is unknown.[1]

4.1.1 Sequential Halving

Sequential Halving was introduced by Karnin et al. (2013). The algorithm is simple: the total sample budget is divided evenly among $\log_2 n$ rounds. In each round arms are sampled uniformly and the worst half of arms in terms of their empirical means are discarded before the next round.

Sequential Halving obtains error control on the order of $\exp\left(-\frac{T}{\mathcal{H}_2 \log_2 n}\right)$ with no prior knowledge of $\mathcal{H}$ or the gaps $(\Delta_i)_{i=2}^n$, and where $\mathcal{H}_2 \coloneqq \max_{i \in [n]} i\Delta_i^{-2}$. We note that $\mathcal{H}_2$ is

[1] Algorithm 7 requires full knowledge of the individual gaps Δ_i, however Audibert, Sebastien Bubeck, and Remi Munos (2010) introduced UCB-E, an algorithm that achieves BAI similar performance via a UCB strategy using only the value of $\mathcal{H}$ to set the exploration bonus for arms.

Require: Sample budget $T \geq n \log_2 n$

$\mathcal{S}_0 \leftarrow [n]$

for round $\ell \in [0 : \log_2 n - 1]$ **do**

  Sample arms in $\mathcal{S}_\ell$ uniformly with $t_\ell = \dfrac{T}{|\mathcal{S}_\ell| \log_2 n}$ samples each

  Set $\mathcal{S}_{\ell+1}$ to be the top half of arms in $\mathcal{S}_\ell$, sorted in descending order by their sample means in round ℓ, $\hat{\mu}_i^\ell$

end for

Return $\psi_T = \mathcal{S}_{\log_2 n}$

Algorithm 8: Sequential Halving (Karnin et al., 2013)

closely related to $\mathcal{H}$:

$$\begin{aligned}\mathcal{H} &= \sum_{i=2}^{n} \frac{1}{\Delta_i^{-2}} = \sum_{i=2}^{n} \frac{i\Delta_i^{-2}}{i} \\ &\leq \sum_{i=2}^{n} \frac{\left(\max_{j \cdot \in [n]} j\Delta_j^{-2}\right)}{i} = \mathcal{H}_2 \cdot \sum_{i=2}^{n} 1/i \leq \mathcal{H}_2 \cdot \log n,\end{aligned}$$

with equality if and only if $\Delta_i = \sqrt{c_1/i}$ for some constant $c_1 > 0$. In the other direction the definition of $\mathcal{H}_2$ implies the existence of some $j \in [2 : n]$ such that $\mathcal{H}_2 = j\Delta_j^{-2}$. From this it follows that $\mathcal{H}_2 = j \cdot \Delta_j^{-2} = \sum_{i=2}^{j} \Delta_j^{-2} \leq \sum_{i=2}^{n} \Delta_i^{-2} = H$, with equality between $\mathcal{H}_2$ and $\mathcal{H}$ if and only if for all $i \geq 2, \Delta_i = c_2$ for some constant $c_2 > 0$ giving us a tight relationship

$$\mathcal{H}_2 \leq \mathcal{H} \leq \log n \cdot \mathcal{H}_2. \tag{4.3}$$

Theorem 4.4 (Karnin et al. (2013, Theorem 4.1))**.** *Running Algorithm 8 (*SH=(π, ψ)*) on an n-armed subgaussian instance ν with budget T yields*

$$\mathbb{P}_{\nu,\pi}\left(\psi_T \neq 1\right) \leq 3 \log_2 n \cdot \exp\left(-\frac{T}{8\mathcal{H}_2 \log_2 n}\right).$$

From Eq. (4.3) we see that on instances with uniform gaps the resulting sample complexity is a factor of $\log_2 n$ worse than the guarantee of uniform exploration, but when the gap structure resembles the case where $\mathcal{H} = \mathcal{H} \log_2 n$, we see that Theorem 4.4 achieves a bound like $\exp\left(-T/\mathcal{H}\right)$, matching the performance of Algorithm 7 without requiring the gap structure as input.

Proof. Achieving a nearly order-optimal upper bound for Sequential Halving is relatively straightforward. We follow the original analysis from Karnin et al. (2013) here, which provides the following bound on the probability that the unique optimal arm is eliminated in round ℓ:

Lemma 4.5.

$$\mathbb{P}\left(1 \notin \mathcal{S}_{\ell+1} | 1 \in \mathcal{S}_\ell\right) \leq 3 \exp\left(-\frac{T}{8 \log_2 n} \cdot \frac{\Delta_{i_\ell}^2}{i_\ell}\right),$$

where we define $i_\ell \coloneqq n/2^{\ell+2}$.

Proof. Our goal is to show that in round ℓ, fewer than $1/3$ of the bottom $3/4$ of arms, when sorting arms in descending order of their true mean values, have sample means that exceed that of the best arm. It follows between the union of the top $1/4$ of arms and the bottom $3/4$ there are not enough arms with sample means larger than the best arm to prevent it from making it to round $\ell + 1$.

Letting $\mathcal{S}'_\ell$ denote the true bottom $3/4$ of arms – i.e. according to their *true* means – of $\mathcal{S}_\ell$, we bound the expected number of arms in this set with an empirical mean larger than the best arm in round ℓ, N_ℓ. On the event $\{1 \in \mathcal{S}_\ell\}$, almost surely,

$$\begin{aligned}
\mathbb{E}\left[N_\ell \mid \mathcal{S}_\ell\right] &\leq \sum_{i \in \mathcal{S}'_\ell} \mathbb{P}\left(\hat{\mu}_i \geq \hat{\mu}_1 \mid \mathcal{S}_\ell\right) \\
&\leq \sum_{i \in \mathcal{S}'_\ell} \exp\left(-\frac{\Delta_i^2 t_\ell}{2}\right) \\
&\leq \sum_{i \in \mathcal{S}'_\ell} \exp\left(-\frac{T}{2 \log_2 n} \frac{2^\ell \Delta_i^2}{n}\right) \\
&\leq |\mathcal{S}'_\ell| \max_{i \in \mathcal{S}'_\ell} \exp\left(-\frac{T}{8 \log_2 n} \frac{\Delta_i^2}{i_\ell}\right) && \text{(recalling } i_\ell = n/2^{\ell+2}\text{)} \\
&\leq \frac{3\,|\mathcal{S}_\ell|}{4} \exp\left(-\frac{T}{8 \log_2 n} \frac{\Delta_{i_\ell}^2}{i_\ell}\right). && \text{(by definition of } \mathcal{S}'_\ell\text{)}
\end{aligned}$$

By Markov's inequality it follows that

$$\mathbb{P}\left(N_\ell > 1/4 \cdot |S_\ell| \mid \mathcal{S}_\ell\right) \leq 3\,|\mathcal{S}_\ell| \exp\left(-\frac{T}{8 \log_2 n} \frac{\Delta_{i_\ell}^2}{i_\ell}\right).$$

□

Finally making use of a union bound over the possibility that we eliminate the best arm in

any of the rounds we have

$$\begin{aligned}
\mathbb{P}_{\text{SH},\nu}\left(\psi_T \neq 1\right) &= \mathbb{P}\left(\exists \ell \in [\log_2 n] : 1 \notin \mathcal{S}_\ell\right) \\
&\leq \sum_{\ell=1}^{\log_2 n} \mathbb{P}\left(1 \notin \mathcal{S}_{\ell+1} | 1 \in \mathcal{S}_\ell\right) \\
&\leq 3 \log_2 n \cdot \max_{\ell \in [\log_2 n]} \exp\left(-\frac{T}{8 \log_2 n} \frac{\Delta_{i_\ell}^2}{i_\ell}\right) \\
&\leq 3 \log_2 n \cdot \max_{i \in [n]} \exp\left(-\frac{T}{8 \log_2 n} \frac{\Delta_i^2}{i}\right) \\
&= 3 \log_2 n \cdot \exp\left(-\frac{T}{8 \mathcal{H}_2 \log_2 n}\right).
\end{aligned}$$

□

We will provide another analysis of Sequential Halving in Section 4.3 when we look at returning ε-good arms with high probability.

4.1.2 The Complexity of Best Arm Identification with a Fixed Budget

Given the apparent similarities of the fixed confidence and fixed budget problem settings it is tempting to assume that the lower bounds from the former translate to the latter. When the learner is given $\mathcal{H}$ for the current instance then this intuition is correct: if we have a fixed-budget algorithm that achieves BAI error of magnitude $\exp\left(-T/\mathcal{H}\right)$, then if we know the value of $\mathcal{H}$ we can construct an (ε, δ)-PAC algorithm with a deterministic stopping time T (simply taking $T = \mathcal{H} \cdot \log(1/\delta)$); by way of contradiction PAC lower bounds apply to the fixed budget setting when $\mathcal{H}$ is known in advance. As we discussed earlier, FB-SE (Algorithm 7) (if the gaps are known) and UCB-E (Audibert, Sebastien Bubeck, and Remi Munos, 2010) achieve the instance-dependent sample complexity $T \sim \mathcal{H} \log(1/\delta)$ in this special case. If this were the optimal sample complexity in the general case of unknown gaps then this would imply an extra $\log n$ factor in the exponent of the performance of the algorithms of Audibert, Sebastien Bubeck, and Remi Munos (2010) and Karnin et al. (2013) which achieve a rate of $\exp(-T/(\mathcal{H}_2 \log_2 n))$. It was not until the work of Garivier and Kaufmann (2016) that it was demonstrated that these algorithms actually achieve the near-optimal exponent for the fixed budget setting. Specifically it was shown that if an algorithm has no knowledge of $\mathcal{H}$ then there exist instances on which an error of order $\exp\left(-T/\mathcal{H} \log n\right)$ is unavoidable.

The lower bound considers classes of n-armed Bernoulli instances parameterized by a, $\mathbb{B}_a^n = \{\nu \in \mathcal{E}_{\mathcal{B}}^n : \mathcal{H}(\nu) \leq a\}$ where $\mathcal{E}_{\mathcal{B}}^n$ is the set of n-armed instances with $\nu_i = \mathcal{B}(\mu_i), \mu_i \in (0,1)$ for $i \in [n]$.

Theorem 4.6 (Garivier and Kaufmann (2016, Theorem 1))**.** *Let $n > 2$, $a > 11n^2$. For any instance $\nu \in \mathbb{B}_a^n$ let $\nu^\star$ denote the arm with the highest mean on ν, where we assume that a unique optimal arm exists. If $T \geq a^2(4\log(6Tn))/60^2$ then for any $\mathsf{Alg} = (\pi, \psi)$ it holds that there exists $\nu \in \mathbb{B}_a^n$ such that*

$$\mathbb{P}_{\nu,\pi}\left(\psi_T \neq \nu^\star\right) \geq \frac{1}{6}\exp\left(-400\frac{T}{\mathcal{H}(\nu)\log n}\right).$$

This theorem demonstrates that, unlike in the fixed confidence setting, algorithms potentially have to pay a factor of $\log n$ in sample complexity in order to adapt to unknown values of $\mathcal{H}$ in the worst case. As mentioned by the authors, at first glance this lower bound would seem to contradict the upper bound in Theorem 4.4 on instances where $\mathcal{H}(\nu) = \mathcal{H}_2(\nu)\log n$. However, this is not the case as Theorem 4.6 only shows that there *is* an instance ν for which the lower bound holds, and so we can expect that on the instance in question we have $\mathcal{H}(\nu) \approxeq \mathcal{H}_2(\nu)$, avoiding a contradiction.

4.2 Simple Regret

We began our discussion of fixed confidence pure exploration with results for the (ε, δ)-PAC setting. There are at least two related objectives in the fixed budget setting. One objective is to bound the probability that an algorithm returns an ε-bad arm as a function of ε and the sample budget T. Alternatively one can look at the rate at which the expected sub-optimality of the returned arm decreases with the size of the sample budget T.

We will begin with the latter metric which we refer to as the 'simple-regret', a term coined by Sébastien Bubeck, Rémi Munos, and Stoltz (2009) in contrast with the cumulative regret studied in the usual bandit setting. As usual we will use uniform exploration as a starting point. Lemma 4.7 follows identically to Lemma 4.1 and on its own adds little insight to this problem.

Lemma 4.7. *Uniformly drawing samples from arms and selecting the empirical best arm (=(π, ψ)) on instance ν with a sampling budget of $T \geq n$ gives*

$$\mathbb{E}_{\nu,\mathrm{UE}}\left[\Delta_{\psi_T}\right] \leq \sum_{i=2}^{n} \Delta_i \exp\left(-\frac{\Delta_i^2 T}{4n}\right)$$

Interestingly the simple regret objective provides an alternative avenue to explore bounds that are known as distribution-free, or minimax upper bounds. Such bounds control the simple regret solely in terms of n and T, regardless of the arms of the instance ν. This analysis was not useful in the BAI setting as the mis-identification probability approaches one on instances where the smallest gap approaches zero. Lemma 4.7 exploits the fact that as the suboptimality gap of arms gets smaller their shrinking contribution to the simple regret compensates for the corresponding growth in error probability.

Corollary 4.7.1 (Sébastien Bubeck, Rémi Munos, and Stoltz (2009, Corollary 3))**.** *Sampling arms with uniform exploration (UE=(π, ψ)) followed by selecting the arm with the empirically highest mean leads to*

$$\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] \leq 4\sqrt{\frac{n \log n}{T}}.$$

Proof. Plugging in Lemma 4.7 we obtain an upper bound on the simple regret as follows:

$$\begin{aligned}
\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] &= \sum_{i=2}^{n} \Delta_i \mathbb{P}_{\nu,\pi}\left(\psi_T = i\right) \\
&\leq \Delta + \sum_{i:\Delta_i > \Delta} \Delta_i \mathbb{P}_{\nu,\pi}\left(\psi_T = i\right) && \text{(for some } \Delta > 0 \text{ to be chosen later)} \\
&\leq \Delta + \sum_{i:\Delta_i > \Delta} \Delta_i \exp\left(-\frac{\Delta_i^2 T}{4n}\right). && \text{(Hoeffding)}
\end{aligned}$$

Noting that $x \exp(-x^2/c)$ reaches its maximum value of $\sqrt{c}/e$ at $x = \sqrt{c}$, then so long as $\Delta \geq \sqrt{\frac{4n}{T}}$ we have

$$\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] \leq \Delta + n\Delta \exp\left(-\frac{\Delta^2 T}{4n}\right),$$

and with the choice of $\Delta = \sqrt{\frac{4n \log n}{T}}$ we have

$$\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] \leq 4\sqrt{\frac{n \log n}{T}}.$$

□

Conveniently, we can make use of distribution-free bounds from the cumulative regret setting to provide bounds on the simple regret.

Theorem 4.8. *If a sampling policy $\pi = (\pi_t)_{t=1}^T$ achieves a distribution-free cumulative regret (Section 2.3) upper bounded by $C\sqrt{nT}$, then we can use this sampling policy to construct an algorithm with a simple regret bounded by $C\sqrt{\frac{n}{T}}$ by running π on our instance, and then recommending an arm with probability proportional to the empirical distribution of arm pulls of π during the exploration phase.*

Proof. Consider an arm selection rule as described above. Then we have

$$\begin{aligned}
\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] &= \sum_{i=2}^{n} \Delta_i \mathbb{P}_{\nu,\pi}\left(\psi_T = i\right) \\
&= \frac{1}{T}\sum_{i=2}^{n} \Delta_i \mathbb{E}_{\nu,\pi}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t = i\}\right] \\
&\qquad\qquad \text{(selecting an arm according to the proportion of samples)} \\
&= \frac{\mathbb{E}_{\nu,\pi}\left[R_T\right]}{T} \\
&\leq C\sqrt{\frac{n}{T}}, \qquad\qquad \text{(by the cumulative regret guarantee for } \pi)
\end{aligned}$$

where in the second last line R_T is the cumulative regret up to round T. □

MOSS (Audibert and Sebastien Bubeck, 2009) is an algorithm that achieves a distribution-free cumulative regret upper bounded by $49\sqrt{nT}$, and so Theorem 4.8 implies that there exists an algorithm that achieves a simple regret of order $\sqrt{\frac{n}{T}}$, albeit this guarantee holds for the slightly more restrictive assumption of arms with distributions bounded in $[0, 1]$ almost surely. This result turns out be optimal.

Theorem 4.9. *For any pure exploration algorithm $\mathsf{Alg} = (\pi, \psi)$, there exists a unit variance gaussian bandit instance ν on which*

$$\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] \geq \frac{1}{e\sqrt{8}}\sqrt{\frac{n}{T}}.$$

Proof. Let ν_1 be a shifted standard normal bandit instance with means $(\Delta, 0, 0, \ldots, 0)$ for some $\Delta > 0$ to be specified later, and ν_2 be another otherwise identical instance with means

$(\Delta, 2\Delta, 0, 0, \ldots, 0)$. Then for any (π, ψ) we have

$$\begin{aligned}
\max\{\mathbb{E}_{\nu_1,\pi}\left[\Delta_{\psi_T}\right], \mathbb{E}_{\nu_2,\pi}\left[\Delta_{\psi_T}\right]\} &\geq \frac{\mathbb{E}_{\nu_1,\pi}\left[\Delta_{\psi_T}\right] + \mathbb{E}_{\nu_2,\pi}\left[\Delta_{\psi_T}\right]}{2} \\
&\geq \frac{\Delta}{2}\left[\mathbb{P}_{\nu_1,\pi}\left(\psi_T \neq 1\right) + \mathbb{P}_{\nu_2,\pi}\left(\psi_T = 1\right)\right] \\
&\geq \frac{\Delta}{4}\exp\left(-\mathrm{KL}\left(\mathbb{P}_{\nu_1,\pi}, \mathbb{P}_{\nu_2,\pi}\right)\right) \\
&\qquad \text{(Bretagnolle-Huber's inequality)} \\
&\geq \frac{\Delta}{4}\exp\left(-\frac{\Delta^2 T}{2n}\right) \\
&\qquad \text{(divergence decomposition lemma Lemma 2.5)} \\
&= \frac{1}{e\sqrt{8}}\sqrt{\frac{n}{T}}. \qquad \text{(choosing } \Delta = \sqrt{\frac{2n}{T}}\text{)}
\end{aligned}$$

□

A similar lower bound holds for Bernoulli instances with means bounded away from $\{0, 1\}$, matching MOSS' result for arm distributions over $[0, 1]$.

We may wonder whether or not the earlier bound in Corollary 4.7.1 is tight – perhaps some smarter arm recommendation scheme other than selecting the empirically best arm could allow for a minimax optimal simple regret using uniform exploration. Theorem 4.10 shows this is not the case: an adaptive sampling strategy is necessary to achieve the optimal minimax regret. The proof relies on the notion of a symmetric algorithm, which we define below.

Definition 4.2 (Symmetric Pure Exploration Algorithm)**.** A symmetric pure exploration algorithm (π, ψ) satisfies $\mathbb{P}_{\nu,\pi}(\psi_T = i) = \mathbb{P}_{\pi,\sigma(\nu)}(\psi_T = \sigma(i))$ where $\sigma \in \Sigma^n$ is an element of the permutation group on sets of size n.

Theorem 4.10. *For any pure exploration algorithm that relies on uniform exploration* $\mathsf{Alg} = (\pi, \psi)$*, where* π *corresponds to uniform exploration and where* ψ *is an arbitrary selection rule such that* Alg *symmetric, there exists a shifted standard normal instance* ν *with* $n \geq 10$ *for which*

$$\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] \geq \frac{1}{9e}\sqrt{\frac{n\log n}{T}}.$$

The proof of Theorem 4.10 can be found in Section C.2.

Having wrapped up the discussion of minimax bounds for simple regret one may understandably come away with the impression that MOSS is an optimal algorithm for

simple regret minimization. As discussed in Section 2.5.2 minimax bounds are not always representative of the typical performance that we expect when we deploy algorithms in the world, and depending on the problem setting can be overly pessimistic and reflect only a hard 'corner' of instance space. Further to this, as mentioned in Chapter 2 it is possible for two algorithms to have the same minimax guarantee but have highly disparate instance-dependent performance. In order to explore more instance dependent results for the simple regret setting Audibert, Sebastien Bubeck, and Remi Munos (2010) noted that for instances with means in $[0, 1]$ the following relationship holds for any algorithm (π, ψ) on an arbitrary instance ν:

$$\Delta_2 \mathbb{P}_{\nu,\pi} \left(\psi_T \neq 1 \right) \leq \mathbb{E}_{\nu,\pi} \left[\Delta_{\psi_T} \right] \leq \mathbb{P}_{\nu,\pi} \left(\psi_T \neq 1 \right). \tag{4.4}$$

From Section 4.1 we know that we can design algorithms for which Eq. (4.4) implies an exponential decrease in T for the simple regret when the gaps of the instance are fixed with respect to T. This relationship was actually the original motivation for studying the BAI objective in the fixed-budget setting, ironically spurring significant progress on the BAI objective for both the fixed budget and fixed-confidence settings in the early 2010's and leaving further analysis of the simple regret performance of algorithms to the wayside until recently.

4.2.1 On the Relationship between the BAI and Simple Regret Objectives

Consider three sample budget regimes for fixed budget pure exploration in bandits: If $T < n$ then we are in the 'infinite-armed' case. Without any prior knowledge of the specific instance that we are interacting with this problem is clearly not solvable; there can always be an optimal arm, or a set of arms with large means that we never observe a sample from. In this setting the BAI-error is not a reasonable objective to minimize as it effectively reduces to solving a needle-in-a-haystack problem on a potentially infinite haystack. On the other hand we can design meaningful algorithms to minimize the simple regret objective on the condition that we are provided with additional structure on the instance. For instance, we may have knowledge about an 'arm reservoir distribution' (Carpentier and Valko, 2015) that characterizes the density of close to optimal arms, or a topology on the arms (Sébastien Bubeck, Rémi Munos, Stoltz, and Szepesvári, 2011) that allows us to reason about how close the means of arms must be to one another depending on their (possibly continuous) arm indices.

On the opposite extreme is the case where the sampling budget $T \gg \mathcal{H}$, where

$\mathcal{H} \coloneqq \sum_{i=2}^{n} 1/\Delta_i^2$ is the characteristic sample complexity we saw in Section 4.1. In this scenario we have enough samples to identify the best arm in the instance with high probability and as such we can upper-bound the simple regret objective by

$$\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] \leq \Delta_n \mathbb{P}_{\nu,\pi}\left(\psi_T \neq 1\right), \tag{4.5}$$

where we have seen algorithms with BAI-error (and therefore simple regret) decaying exponentially with $T/(\mathcal{H}_2 \log_2 n)$. Further, with such a large sampling budget we expect that for any reasonable BAI algorithm we have $\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] \cong \Delta_2 \mathbb{P}_{\nu,\pi}\left[\psi_T \neq 1\right]$, where the smallest gap has replaced the largest gap in Eq. (4.5), since the arm with the smallest gap will likely have the largest contribution to the error in choosing the best arm.

The third sample size regime is sandwiched between the previous two: when we have sufficient sampling budget such that all arms can be sampled, but insufficient budget to identify the best arm in the instance with arbitrarily high confidence, i.e. we consider the 'many-armed bandit' setting defined by a sample budget range $T \cong H$. It is in this case where a non-trivial relationship between the BAI and simple regret objectives can exist. In this regime both the BAI-error and simple regret objectives are meaningful in the sense that neither objective is trivial to optimize. Recall that we have argued that when $T < n$ then simply recommending an arm at random is an un-improvable strategy for controlling the BAI-error, as we saw in the needle-in-a-haystack example in Section 2.5, and as $T \gg \mathcal{H}$ we see that the two objectives effectively merge. The question remains whether there are problem instances in this intermediate regime where algorithms must be designed with specific consideration as to which objective they seek to minimize. Specifically we are interested in understanding the possibility of 'best-of-both-worlds' algorithms for the objectives, or whether optimal data collection strategies on each objective differ.[2]

Recall that Eq. (4.4) shows that finding the best arm on an instance is sufficient to a identify a good instance (in the sense of simple regret), and naturally any good arm will necessarily be no more optimal than the best arm. The full implication of Eq. (4.4) is that for a fixed instance (i.e. keeping n and the arm means fixed), the simple regret is controlled by the rate of decay of the BAI-error up to the gap between the smallest and largest gaps on the instance.

[2]We focus on data collection strategies here, since if the only change required for each objective is in the final arm selection function then practitioners can run one sampling procedure and then apply two different arm selection rules to optimize each objective.

Consider the BAI-error and simple regret objectives for the case $n = 3$ and $\Delta_2 \ll \Delta_3 = 1$:

$$\mathbb{P}_{\nu,\pi}\left(\psi_T \neq 1\right) = \mathbb{P}_{\nu,\pi}\left(\psi_T = 2\right) + \mathbb{P}_{\nu,\pi}\left(\psi_T = 3\right), \tag{4.6}$$

$$\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] = \Delta_2 \mathbb{P}_{\nu,\pi}\left(\psi_T = 2\right) + \mathbb{P}_{\nu,\pi}\left(\psi_T = 3\right). \tag{4.7}$$

Intuitively, optimizing Eq. (4.6) requires us to balance the two terms on the right hand side. If all arms are 1-subgaussian then doing so roughly equates to sampling the sub-optimal arms until their confidence intervals diverge from that of the best arm, and that requires approximately $1/\Delta_2^2$ samples for arm 2, and a constant number of samples for arm 3. In contrast, balancing the terms to optimize Eq. (4.7) is a different matter. Because $\Delta_2 \ll 1$, the simple regret places less importance on making the mistake $\psi_T = 2$ than $\psi_T = 3$, and we expect that more samples on arm 3 would be needed to optimize Eq. (4.7) than for Eq. (4.6). This statement alone does not show that the BAI-error and simple regret objectives are misaligned, sinc the term that is more heavily weighted in the simple regret corresponds to recommending a highly sub-optimal arms, and arms that are more sub-optimal are in turn easier to identify and discard by algorithms, as we saw in the proof of Lemma 4.7.

To address this consider an instance where the majority of arms have a constant suboptimality gap. Specifically let ν have $n > 3$ arms with $\Delta_2 \ll \Delta_3 = \Delta_4 = \cdots = \Delta_n = 1$. Then

$$\mathbb{P}_{\nu,\pi}\left(\psi_T \neq 1\right) = \mathbb{P}_{\nu,\pi}\left(\psi_T = 2\right) + \mathbb{P}_{\nu,\pi}\left(\psi_T > 2\right), \tag{4.8}$$

$$\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] = \Delta_2 \mathbb{P}_{\nu,\pi}\left(\psi_T = 2\right) + \mathbb{P}_{\nu,\pi}\left(\psi_T > 2\right). \tag{4.9}$$

Now even though the weighting of the second probability in Eq. (4.9) is still a constant this probability will require more than a constant number of samples to in order to be bound by a constant, since the learner needs to control the probability that *any* of the sub-optimal arms are chosen. This allows for the possibility of a situation where $\mathbb{P}_{\nu,\pi}\left(\psi_T = 2\right)$ is the dominant term in Eq. (4.8) whereas $\mathbb{P}_{\nu,\pi}\left(\psi_T > 2\right)$ dominates Eq. (4.9), necessitating different data collection strategies to minimize each respective loss. For instance, suppose that the gap structure of ν is such that $\Delta_2 = \sqrt{\log\log n/T}$, and $\inf_{(\pi,\psi)} \mathbb{P}_{\nu,\pi}\left(\psi_T \neq 1\right) = \Theta\left(1/\log n\right)$. With $T \geq n$, Eq. (4.4) shows

$$\sqrt{\frac{\log\log n}{T}}\frac{1}{\log n} \lesssim \mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] \leq \frac{1}{\log n}. \tag{4.10}$$

Clearly the upper and lower bounds are of different orders in $T \geq n$ revealing a gap in which it is no longer clear that optimal BAI control yields optimal control of the simple regret. The next section attempts to formalize and explore this idea.

Optimal Data Collection Oracles for the BAI and Simple Regret Objectives

Consider a modified pure exploration setting with a fixed sample budget of T in which we have a non-stochastic 'control' arm that always returns a value of 1, a 'good' shifted standard normal arm with suboptimality gap $\sqrt{\frac{\log\log n}{T}}$ and n 'bad' shifted standard normal arms all with a suboptimality gap of 1. We consider the problem of optimizing the batch allocation of the T samples to arms by an oracle that knows the mean of each arm in advance, so as to optimize the pure exploration performance of selecting the empirically best arm ψ_T when the sample budget is $T = C \cdot n \log n$ for some $C > 0$ to be chosen later. This choice of T is close to $\mathcal{H} = \frac{T}{\log\log n}\left(1 + \frac{\log\log n}{C \log n}\right)$ and so fits into the intermediate sample regime from the earlier discussion. By taking the approach of scaling the budget T with the size of the instance we obtain similar results to those in high-dimensional statistics in that the natural quantity for scaling is T/d, where d is the dimension of the problem setting and T is the number of samples (e.g.Wainwright (2019)).

Definition 4.3. Oracle Allocation Game with Control Arm

Require: Stochastic arm gaps $\boldsymbol{\Delta} = (\Delta_i)_{i=2}^{n+2}$, sample budget $T \in \mathbb{N}$

1: The oracle $\mathcal{O}$ chooses an allocation vector $\alpha = (\alpha_i)_{i=2}^{n+2} \in \mathcal{M}_{n+2}$ that can depend on the arm gaps $\boldsymbol{\Delta}$ and T
2: $\alpha_i T$ i.i.d samples are collected from each arm $i \in [2 : n+2]$ to obtain arm sample means $(\hat{\mu}_i)_{i\in[2:n+2]}$
3: Arm $\psi_T \in \arg\max_{i\in[n+2]} \hat{\mu}_i$ is selected, where $\hat{\mu}_1 := 1$
4: The oracle receives loss $\ell(\psi_T, \boldsymbol{\Delta})$

We evaluate the quality of an oracle $\mathcal{O}$ by the BAI error $\mathbb{P}_{\mathcal{O}}(\psi_T = 1)$ or the simple regret $\mathbb{E}_{\mathcal{O}}[\Delta_{\psi_T}]$ where $\mathbb{P}_{\mathcal{O}}$ denotes the probability measure considering the randomness of the samples observed in drawing samples according to allocation oracle $\mathcal{O}$.

By considering such a constrained problem setting we can make strong statements on the optimality of allocation oracles. We will consider a parameterized family of allocation oracles $\mathcal{O}_\gamma$, where $\gamma \in (0, 1)$ specifies the oracle allocation vector $\alpha = \alpha(\gamma) = (\gamma, (1-\gamma)/n, (1-\gamma)/n, \ldots, (1-\gamma)/n)$ with γ proportion of samples on arm 2, and the

remaining $1-\gamma$ proportion spread over the n mean zero arms. Lemma 4.11 bounds the error of this family of oracles.

Lemma 4.11 (Upper Bounds on Oracle Error Probabilities)**.** *For oracle $\mathcal{O}_\gamma$ and sample budget $T = Cn\log n$*

$$\mathbb{P}_\gamma(\psi_T = 2) \leq (\log n)^{-\gamma/2}$$

$$\mathbb{P}_\gamma(\psi_T > 2) \leq n^{1-C(1-\gamma)/2}$$

where $\mathbb{P}_\gamma$ denotes the probability measure induced by playing oracle $\mathcal{O}_\gamma$ on the instance.

Proof. Using $\hat{\mu}_1 \coloneqq 1$,

$$\begin{aligned}
\mathbb{P}_\gamma(\psi_T = 2) &\leq \mathbb{P}_\gamma(\hat{\mu}_2 \geq 1) \\
&\leq \mathbb{P}_\gamma\left(\hat{\mu}_2 - \mu_2 \geq \sqrt{\frac{\log\log n}{T}}\right) \\
&\leq \exp\left(-\frac{\gamma T \log\log n}{2T}\right) && \text{(Hoeffding)} \\
&= (\log n)^{-\gamma/2}.
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}_\gamma(\psi_T > 2) &\leq \mathbb{P}_\gamma(\exists i \in [3 : n+2] : \hat{\mu}_i \geq 1) \\
&\leq n \cdot \mathbb{P}_\gamma(\hat{\mu}_3 \geq 1) && \text{(union bound, } (\hat{\mu}_i)_{i\geq 3} \text{ are identically distributed)} \\
&\leq n \cdot \exp\left(-\frac{(1-\gamma)T}{2n}\right) && \text{(Hoeffding)} \\
&\leq n^{1-C(1-\gamma)/2}. && (T = Cn\log n)
\end{aligned}$$

□

A corresponding lower bound applies for a general family of allocation oracles and shows that $\mathcal{O}_\gamma$ contains an asymptotically optimal oracle for both objectives (ignoring lower order terms).

Lemma 4.12 (General Lower Bound on Oracle Game Errors)**.** *For a family of oracle strategies $\tilde{\mathcal{O}}_\beta$ that assign βT samples to the arm with the smallest gap (note that $\tilde{\mathcal{O}}_\beta$ is a larger family of oracle allocations than $\mathcal{O}_\gamma$ since there is complete freedom in the allocation across the*

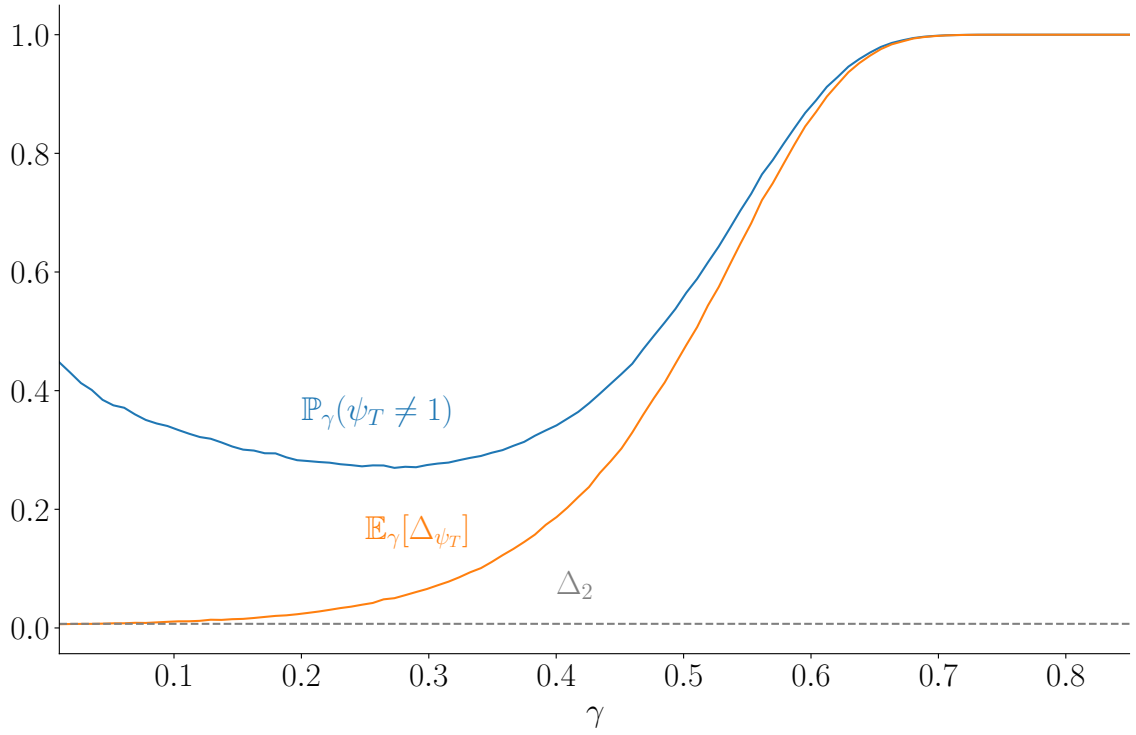

Figure 4.1: Pure exploration performance of the $\mathcal{O}_\gamma$ oracle allocation family for $n = 1000$, averaged over 100,000 replications with $T = 3n \log n$. With such a small sample budget the probability of selecting a mean zero arm dominates the simple regret and vanishingly small values of γ are optimal. In contrast, due to the small gap between the top two arms a constant fraction of samples needs to be allocated to the competitor arm for optimal BAI error (approaching 1/3 from below in the limit as n increases).

remaining n stochastic arms) and sample budget $T = Cn\log n$. Writing the fractional allocation to the i^{th} 'bad' arm as β_i for $i \geq 3$ we have

$$\begin{aligned}\mathbb{P}_{\tilde{O}_\beta}(\psi_T = 2) &\geq \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\beta\log\log n + 4}}\left(\log n\right)^{-\frac{\beta}{2}}\left[1 - \frac{1}{\sqrt{2\pi}}\sum_{i=3}^{n+2}\frac{e^{-C\beta_i n\log(n)/2}}{\sqrt{\beta_i C n\log n}}\right] \\ &\geq \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\beta\log\log n + 4}}\left(\log n\right)^{-\frac{\beta}{2}}\left[1 - \frac{1}{\sqrt{2\pi}}\frac{n^{1-\frac{C}{2}(1-\beta)}}{\sqrt{(1-\beta)C\log n}}\right], \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(Jensen's inequality)}\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}_{\tilde{O}_\beta}(\psi_T > 2) &\geq \sum_{i=3}^{n+2}\frac{1}{\sqrt{8\pi}}\frac{e^{-C\beta_i n\log(n)/2}}{\sqrt{\beta_i C n\log n + 4}} \\ &\geq \frac{1}{\sqrt{8\pi}}\frac{n^{1-\frac{C}{2}(1-\beta)}}{\sqrt{(1-\beta)C\log n + 4}} \qquad\qquad \text{(Jensen's inequality)}\end{aligned}$$

where in the last line of each lower bound we make use of the constraint that $\sum_{i=3}^{n+2}\beta_i = (1-\beta)$ along with the convexity in x of $\frac{e^{-x}}{\sqrt{ax+b}}$ for $a, b > 0$.

Note that for both terms the lower bound is most relaxed for an oracle allocation with equal allocation across all mean zero arms, implying that without loss of generality we can consider optimal oracles among this more constrained family.

Proof. The proof relies on a bound on the tails of normal random variables:

Lemma 4.13 (Milton, Abramowitz (1974, Formula 7.1.13))**.**

$$\frac{1}{x + \sqrt{x^2 + 4}} < e^{x^2}\int_x^\infty e^{-t^2/2}dt < \frac{1}{x + \sqrt{x^2 + 8/\pi}}.$$

Recall that for a shifted standard normal variable Z we have

$$\mathbb{P}\left(Z \geq x\right) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt, \tag{4.11}$$

and so Lemma 4.13 provides us with an anti-concentration result for our sample means (which

are normal variables in this setting):

$$\begin{aligned}
\mathbb{P}_{\check{\mathcal{O}}_\beta}(\psi_T = 2) &\geq \mathbb{P}_{\check{O}_\beta}\left(\hat{\mu}_2 \geq 1, \forall i \in [3 : n+2] : \hat{\mu}_i < 1\right) \\
&= \mathbb{P}_{\check{O}_\beta}\left(\hat{\mu}_2 \geq 1\right) \cdot \mathbb{P}_{\check{\mathcal{O}}_\beta}\left(\forall i \in [3 : n+2] : \hat{\mu}_i < 1\right) \\
&\qquad \text{(independence of sample means)} \\
&\geq \mathbb{P}_{\check{\mathcal{O}}_\beta}\left(\hat{\mu}_2 \geq 1\right) \cdot \left[1 - \mathbb{P}_{\check{\mathcal{O}}_\beta}\left(\exists i \in [3 : n+2] : \hat{\mu}_i \geq 1\right)\right] \\
&\geq \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\beta \log\log n + 4}} \left(\log n\right)^{-\frac{\beta}{2}} \left[1 - \frac{1}{\sqrt{2\pi}} \sum_{i=3}^{n+2} \frac{e^{-C\beta_i n \log(n)/2}}{\sqrt{\beta_i C n \log n}}\right], \\
&\qquad \text{(Lemma 4.13)}
\end{aligned}$$

furthermore,

$$\begin{aligned}
\mathbb{P}_{\check{\mathcal{O}}_\beta}(\psi_T > 2) &\geq \mathbb{P}_{\check{\mathcal{O}}_\beta}\left(\exists i \in [3 : n+2] : \hat{\mu}_i \geq 1,\ \hat{\mu}_2 < \mu_2\right) \\
&= \mathbb{P}_{\check{\mathcal{O}}_\beta}\left(\hat{\mu}_2 < \mu_2\right) \cdot \mathbb{P}_{\check{\mathcal{O}}_\beta}\left(\exists i \in [3 : n+2] : \hat{\mu}_i \geq 1\right) \\
&\qquad \text{(independence of sample means)} \\
&= \frac{1}{2} \cdot \sum_{i=3}^{n+2} \mathbb{P}_{\check{\mathcal{O}}_\beta}\left(\hat{\mu}_i \geq 1\right) \\
&\geq \frac{1}{2} \frac{1}{\sqrt{2\pi}} \sum_{i=3}^{n+2} \frac{e^{-C\beta_i n \log(n)/2}}{\sqrt{\beta_i C n \log n + 4}}., \qquad \text{(Lemma 4.13)}
\end{aligned}$$

□

These results appear complicated but for our purposes it will suffice to consider the order of these bounds in n for specific values ranges of values for γ, C. From Lemma 4.12 we see that in order to achieve a BAI error and simple regret that decreases with at least a poly-logarithmic rate in n we need $\gamma \leq 1 - 2/C$ in order to control the $n^{1-C(1-\gamma)/2}$ lower bound for $\mathbb{P}_\gamma(\psi_T > 2)$ and $\gamma > 0$ to control $\mathbb{P}_\gamma(\psi_T = 2)$. Restricting our attention to $\gamma < 1 - 2/C$, the leading order of growth for the lower bound on $\mathbb{P}_\gamma(\psi_T = 2)$ becomes $(\log n)^{-\gamma/2}$ (Lemma 4.11). It follows that when $\gamma < 1 - 2/C$ there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$

$$\left(c_1 \frac{(\log n)^{-\frac{\gamma}{2}}}{\sqrt{\log\log n}} \leq\right) \mathbb{P}_\gamma(\psi_T = 2) \leq \mathbb{P}_\gamma(\psi_T \neq 1) \leq c_2 \cdot (\log n)^{-\gamma/2} \tag{4.12}$$

for constants $c_1, c_2 > 0$. This implies that minimizing the asymptotic BAI error amounts to maximizing the samples on arm 2, setting $\gamma = 1 - 2/C + \varepsilon$ for $\varepsilon > 0$ some arbitrarily small positive constant. Now considering the simple regret we observe that if $C \leq 3$ then $1 - C(1-\gamma)/2 \geq -1/2$ and so $\mathbb{P}_\gamma\left(\psi_T > 2\right) = \Omega\left(n^{1-C(1-\gamma)/2}\right)$ decreases at a slower rate

than $\Delta_2 \cdot \mathbb{P}_\gamma(\psi_T = 2) \leq \sqrt{\frac{\log\log n}{Cn\log n}} \cdot (\log n)^{-\gamma/2}$. This means that there exists another $n_0 \in \mathbb{N}$ for which if $n \geq n_0$

$$\left(c_1 \cdot \frac{n^{1-\frac{C}{2}(1-\gamma)}}{\sqrt{\log n}} \leq\right) \mathbb{P}_\gamma(\psi_T > 2) \leq \mathbb{E}_\gamma[\Delta_{\psi_T}] \leq c_2 \cdot n^{1-\frac{C}{2}(1-\gamma)} \tag{4.13}$$

for a new set of $c_1, c_2 > 0$. It follows that if we take $C = 3$, then the left hand side of the display above is at least of order $\Omega(1/\sqrt{n})$ (ignoring poly-logarithmic terms in n), and can only be approached by the upper bound for $\gamma = \varepsilon$ for some arbitrarily small positive constant $\varepsilon > 0$. Note that this is what we observe in Fig. 4.1, which shows Monte-Carlo estimates for the losses of the $\mathcal{O}_\gamma$ oracle family on an instance with $n = 1000$ and $C = 3$.

To summarize, upper and lower bounds for oracle allocation families show that as n increases, the optimal value of γ, the proportion of samples allocated to arm 2, which has an vanishing suboptimality gap, approaches $1/3$ from below for the BAI objective, and zero from above for the simple regret. The intuition behind this result is that as n – the number of bad arms – increases, the suboptimality of gap of arm 2 *decreases* by design of Δ_2. On such an instance optimizing the BAI objective amounts to uniformly sampling the n bad arms just enough that the probability of mistakenly choosing one of these bad arms is guaranteed to be dominated by the probability of selecting arm 2, which requires a constant proportion of the total sampling budget thanks to the choice $T = Cn\log n$. From there the rest of the samples are dedicated to reducing the error on arm 2. For the simple regret the regret contribution from arm 2 is of order $1/\sqrt{n}$ due to the small suboptimality gap of arm 2, and so when the sampling budget is small ($C \leq 3$) the dominant regret term is coming from mistakes made on bad arms and so (almost) all samples are placed on these bad arms instead.

Finally we have arrived at a result that touches on the initial motivation for this analysis. Let us consider the simple regret performance of the optimal BAI oracle, and vice versa. By the analysis above we see that using the optimal value of $\gamma = 1/3 - \varepsilon$ for BAI, the simple regret is lower bounded by the probability of choosing a mean zero arm which scales like $\tilde{\Theta}\left(n^{-\varepsilon/2}\right)$ where we ignore polylog factors in n. Going the other direction, using the optimal allocation for the simple regret, $\gamma = \varepsilon$ we see that the BAI error is dominated by the probability of selecting arm 2, which scales like $\tilde{\Theta}\left((\log n)^{-\varepsilon/2}\right)$ where we ignore $\log\log n$ factors. Given that $\varepsilon > 0$ is arbitrarily small we see that the optimal oracle for one objective is incompatible with the other, resulting in an effectively constant loss on one and the optimal rate on the other.

Once again we note that these results are only possible for a narrow range of sampling budgets.

The results in this section are far from general due to the worst-case nature of the construction as well as the restriction to considering oracle allocations, however to the best of our knowledge at the time of writing they consist of the only proof and characterization of a divergence between optimal data collection policies for the BAI and simple regret objectives. An apparent mismatch between the simple regret and BAI performance of allocation oracles was noted incidentally by Russo (2020) but not explored further. One reason for this is likely that the sample budget regime where this effect is possible makes for challenging analysis, since it is necessarily when not there are not enough samples available for the sample means of all arms to concentrate. To address this issue the approach in this section was to consider asymptotic results as the number of arms and sample budget increase with a fixed relationship between these quantities. Recent work has begun to analyse the convergence of sampling distributions for various algorithms in the limit of small gaps, e.g. Kalvit and Zeevi (2021). This direction of research may allow the results in this section to be extended to a more general class of algorithms and problem settings.

4.3 Faster Rates for Identifying Good Arms

The results discussed in this subsection are based on ongoing collaborative work with Yao Zhao and Kwang-Sung Jun at the University of Arizona.

In Section 3.3 we covered some practical limitations of the (ε, δ)-PAC framework for analyzing algorithms. Namely it was shown that on problems for which $m \leq n$ arms are ε-good the stopping times of PAC algorithms necessarily increase with the number of competitive arms and in fact scale like $\Omega\left(m/\varepsilon^2\right)$. This is at odds with our intuition that when $m > 2$, finding one of the ε-good arms should be easier. We noted in Chapter 3 that the reason for this mis-match is the requirement on PAC learners to verify their recommendation. In order to construct well defined sample complexities with the expected dependence on the number of good arm we had to move away from the standard PAC framework, e.g. studying the unverifiable sample complexity (Definition 3.3) of an algorithm on a given instance. In comparison, the results we are looking for come naturally in the fixed budget setting as it lacks a verification requirement. Due to this we can see that instance-dependent lower bounds for (ε, δ)-PAC arm identification

such as Theorem 3.5 have limited bearing on results in the fixed budget setting:

Recall that for (ε, δ)-PAC arm identification, Theorem 3.5 states that on instance ν any (ε, δ)-PAC algorithm must satisfy $\mathbb{E}_{\nu,\mathsf{Alg}}[\tau] = \Omega\left[\log\left(1/(2.4\delta)\right) \cdot \sum_{i=2}^{n} 1/(\Delta_i^{\varepsilon})^2\right]$ where $\Delta_i \coloneqq \max\{\varepsilon, \Delta_i\}$. Now suppose that (π, ψ) is a fixed budget algorithm with $\mathbb{P}_{\nu,\pi}\left(\Delta_{\psi_T} \geq \varepsilon\right) \leq \exp\left(-T/\mathcal{G}_\nu(\varepsilon)\right)$ for some $\mathcal{G}_\nu(\varepsilon) = o\left(\sum_{i=2}^{n} 1/(\Delta_i^{\varepsilon})^2\right)$. There is no contradiction with Theorem 3.5 since in order to convert (π, ψ) to a fixed confidence algorithm we need to be able to construct a stopping time (sample budget) τ that we can set in advance and without knowing ν. In order to guarantee that $\exp\left(-\tau/\mathcal{G}_\nu(\varepsilon)\right) \leq \delta$ we set $\tau \geq \mathcal{G}_\nu(\varepsilon)\log(1/\delta)$. Since $\mathcal{G}_\nu(\varepsilon)$ depends on the unknown instance ν the best we can do in a general setting is to use a worst-case upper bound $\tau/\log(1/\delta) = n\varepsilon^{-2} \geq \sum_{i=2}^{n} 1/(\Delta_i^{\varepsilon})^2 \geq \mathcal{G}_\nu(\varepsilon)$ which leads to a suboptimal sample complexity $\tau = n\varepsilon^{-2}\log(1/\delta)$, avoiding any contradiction.

Having seen that fixed budget algorithms are not necessarily constrained to sample complexities that grow like $\Omega\left(m\varepsilon^{-2}\right)$ on instances with m ε-good arms we may wonder what the lower bounds are for the sample complexity in this setting. While PAC lower bounds do not apply to the fixed budget setting the lower bounds for (ε, δ)-unverifiable stopping times do.

Recall that Theorem 3.14 tells us that for fixed $\varepsilon > 0, \delta \in (0, 1/16)$ and for any symmetric algorithm (π, ψ), then if on an n-armed 1-subgaussian instance ν an $\mathcal{F}$-adapted stopping time $\mathcal{T}$ satisfies $\mathbb{P}_{\nu,\pi}\left(\exists t \geq \mathcal{T} : \Delta_{\psi_t} \geq \varepsilon\right) \leq \delta$, that is, $\mathcal{T}$ is an (ε, δ)-unverifiable stopping time of Alg with respect to ν, then it also satisfies $\mathbb{E}_{\nu,\pi}[\mathcal{T}] \geq \mathcal{H}_{\text{low}}(\nu, \varepsilon) \coloneqq \frac{1}{64}\left(-\Delta_{m+1}^{-2} + \frac{1}{m}\sum_{i=m+1}^{n} \Delta_i^{-2}\right)$, where m is the number of ε-good arms on ν and where we have made use of the assumption of the symmetry of the algorithm to remove an average over permutations of the instance in our original statement of Theorem 3.14. This theorem lacks a dependence on δ, but still allows us to construct a lower bound on sample complexity.

Theorem 4.14 (A weak sample complexity bound for finding ε-good arms)**.** *Fix $\varepsilon > 0$. Let ν be an arbitrary 1-subgaussian instance. Consider a symmetric fixed budget algorithm (π, ψ) which satisfies $\mathbb{P}_{\nu,\pi}\left(\Delta_{\psi_T} \geq \varepsilon\right) \leq \exp\left(-T_0/\mathcal{G}(\nu, \varepsilon)\right)$ for all $T \geq T_0$, for some $T_0 \leq \mathcal{H}_{\text{low}}(\nu, \varepsilon)$. Then*

$$\mathcal{G}(\nu, \varepsilon) \geq \frac{\mathcal{H}_{\text{low}}(\nu, \varepsilon)}{\log(16)}.$$

Proof. The proof follows by showing the relationship between fixed budget performance guarantees and unverifiable sample complexities and from there following the implications of Theorem 3.14.

Given the said symmetric fixed budget algorithm, for sample budget T we define $\psi_t \coloneqq \psi_T$ for $t \geq T$. For such a modified algorithm and for any n-armed 1-subgaussian instance ν, any $T \geq T_0$ is an $(\varepsilon, \exp(-T/\mathcal{G}(\nu, \varepsilon)))$-unverifiable stopping time. We conclude the proof by way of a contradiction: assume that $\mathcal{G}(\nu, \varepsilon) \geq \mathcal{H}_{\text{low}}(\nu, \varepsilon)/\log(16)$. Then for $T = T_0$ we have

$$\begin{aligned}
\exp\left(-\frac{T_0}{\mathcal{G}(\nu, \varepsilon)}\right) &< \exp\left(-\frac{T_0 \log(16)}{\mathcal{H}_{\text{low}}(\nu, \varepsilon)}\right) && \text{(by assumption)} \\
&\leq \exp\left(-\frac{\mathcal{H}_{\text{low}}(\nu, \varepsilon) \log(16)}{\mathcal{H}_{\text{low}}(\nu, \varepsilon)}\right) && \text{(using } T_0 \leq \mathcal{H}_{\text{low}}(\nu, \varepsilon)\text{)} \\
&= 1/16, && (4.14)
\end{aligned}$$

which contradicts Theorem 3.14, where we recall that we consider a symmetric algorithm such that $\mathbb{P}_{\sigma(\nu),\pi}(\Delta_{\psi_T}) = \mathbb{P}_{\nu,\pi}(\Delta_{\psi_T})$ for all $\sigma \in \Sigma^n$. The result follows by algebra. □

Loosely, Theorem 4.14 says that the error control we can achieve with fixed budget algorithms on an instance ν must either be no stronger than $\exp(-T/\mathcal{H}_{\text{low}}(\nu, \varepsilon))$, or only hold for values of $T \geq \mathcal{H}_{\text{low}}(\nu, \varepsilon)$. In either case this says that the sample complexity of identifying an ε-good arm with probability at least $15/16$ in the fixed budget setting is at least $\mathcal{H}_{\text{low}}(\nu, \varepsilon)$, which for an instance with m arms with mean μ_0 and $n - m$ arms with mean $\mu_0 - \varepsilon$ leads us to aim for error control like

$$\mathbb{P}_{\nu,\pi}(\Delta_{\psi_T}) \leq \exp\left(-m\varepsilon^2 T/n\right).$$

Such a result would actually be un-improvable as one can show that worst-case $(n, m, \varepsilon, \delta)$-PAC lower bounds such as Theorem 3.13 apply to this setting.

Revisiting Sequential Halving

Sequential Halving (SH, Algorithm 8) is a parameter-free pure exploration algorithm that we covered in Section 4.1. The algorithm operates in $\log_2 n$ phases of equal sample budgets, uniformly sampling arms in each phase and only keeping the top performing half of arms for the next phase. It turns out that in addition to obtaining nearly optimal BAI performance SH also achieves error control for ε-good arm identification that is nearly in line with Section 4.3.

To see this we first provide a bound on the error probability for identifying an (m, ε)-good arm, i.e. recommending an arm such that $\Delta_{\psi_T} < \Delta_m + \varepsilon$:

Theorem 4.15. *Suppose that we run Sequential Halving ($= (\pi, \psi)$) on 1-subgaussian instance ν with a budget of $T \geq n \log_2 n$. Then for any $m \leq n$,*

$$\mathbb{P}_{\nu,\pi}\left(\Delta_{\psi_T} \geq \Delta_m + \varepsilon\right) \leq \log_2 n \cdot \exp\left[-C \cdot m \left(\frac{\varepsilon^2 T}{[4(\log_2(2m))^2 \log_2 n] \cdot n} - \log(4e)\right)\right],$$

where $C > 0$ is a universal constant.

It follows that when $T \geq \left[\left(\log(4e) + \frac{\log\log_2 n}{C}\right)(\log_2(2m))^2 \log_2 n\right] \cdot \frac{n}{\varepsilon^2}$, we have

$$\mathbb{P}_{\nu,\pi}\left(\Delta_{\psi_T} \geq \Delta_m + \varepsilon\right) \leq \exp\left(-\frac{C}{4(\log_2(2m))^2 \log_2 n}\left(m\frac{\varepsilon^2 T}{n}\right)\right).$$

Noticing that Theorem 4.15 holds without providing m or ε as input to the algorithm leads us to an accelerated rate of error reduction: Theorem 4.15 implies that for any pair $(\tilde{m}, \tilde{\varepsilon} > 0)$ that satisfies $\Delta_{\tilde{m}} + \tilde{\varepsilon} < \varepsilon$ we have

$$\mathbb{P}_{\nu,\pi}\left(\Delta_{\psi_T} \geq \varepsilon\right) \leq \exp\left(-\tilde{\Theta}\left(\tilde{m}\frac{\tilde{\varepsilon}^2 T}{n}\right)\right)$$

where $\tilde{\Theta}$ hides logarithmic dependence in $n, \tilde{m}$. Notice that for a fixed value of $\tilde{\varepsilon}$ choosing the largest compatible value of $\tilde{m} : \Delta_{\tilde{m}} + \tilde{\varepsilon} \leq \varepsilon$ results in a tighter bound, and alternatively choosing a larger value of $\tilde{\varepsilon} < \varepsilon$, while initially loosening the bound may enable a larger choice of $\tilde{m}$ leading to an overall tighter bound. Fig. 3.3 illustrates the situation. The previous theorem holds for all valid pairs $(\tilde{m}, \tilde{\varepsilon})$ allowing for a tightened bound on the error of identifying an ε-good arm:

Corollary 4.15.1. *Suppose we run Sequential Halving ($= (\pi, \psi)$) on an n-armed 1-subgaussian instance ν with a budget of $T \geq n \log_2 n$. Then for any $\varepsilon > 0$,*

$$\mathbb{P}_{\nu,\pi}\left(\Delta_{\psi_T} \geq \varepsilon\right) \leq \min_{(\tilde{m},\tilde{\varepsilon}):\Delta_{\tilde{m}}+\tilde{\varepsilon}\leq\varepsilon} \log_2 n \cdot \exp\left[-C \cdot \tilde{m}\left(\frac{\tilde{\varepsilon}^2 T}{[4(\log_2(2\tilde{m}))^2 \log_2 n] \cdot n} - \log(4e)\right)\right],$$

where $C > 0$ is a universal constant.

Going back to two-level instances ν with m mean μ_0 arms and $n - m$ arms mean $\mu_0 - \varepsilon$, Corollary 4.15.1 shows that on these instances Sequential Halving is able to return one of the optimal arms after $\tilde{O}\left(m/n \cdot \varepsilon^{-2} \log(1/\delta)\right)$ samples, matching the weak lower bound on the

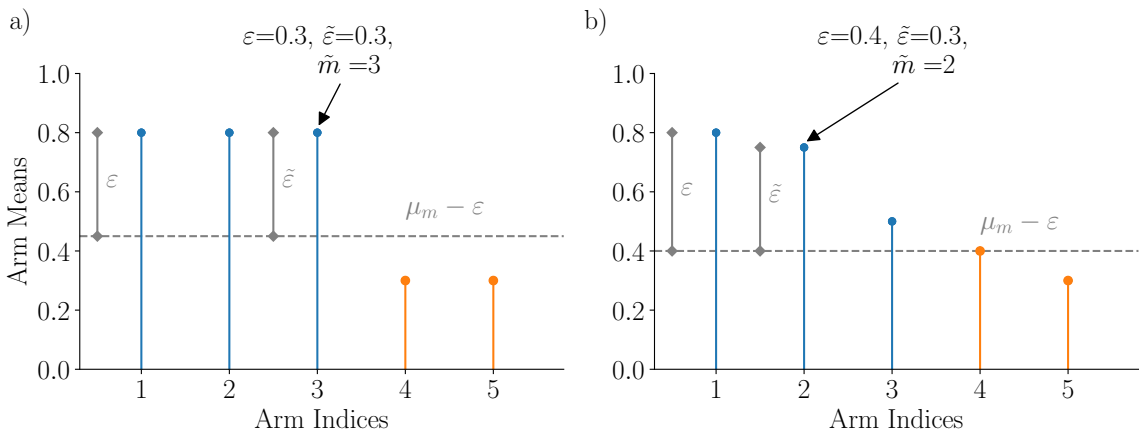

Figure 4.2: The pair $\tilde{m}, \tilde{\varepsilon}$ that minimizes the bound in Corollary 4.15.1 roughly correspond to the pair that maximizes the quantity $\tilde{m} \cdot \tilde{\varepsilon}^2$ while satisfying $\Delta_m + \tilde{\varepsilon} < \varepsilon$. We show examples of such pairs on two different instances. In each of these figures the arms with blue stems are ε-good, and arms with orange stems are ε-bad. The improvement in the sample complexity guarantee from the näive choice of $\tilde{m} = 1, \tilde{\varepsilon} = \varepsilon$ on the instance in a) is $\tilde{m} \cdot \tilde{\varepsilon}^2 = 3\varepsilon^2$ and in b) is $\tilde{m} \cdot \tilde{\varepsilon}^2 = \frac{25}{16}\varepsilon^2$.

sample complexity in Theorem 4.14 up to logarithmic factors. We note that this algorithm can be applied to the $(m, n, \varepsilon, \delta)$-PAC setting from Section 3.3 by setting $\tau = \tilde{O}\left(m/n \cdot \varepsilon^{-2} \log(1/\delta)\right)$, resulting in an algorithm that improves upon the sample complexity of Algorithm 5 by a factor of $\log(1/\delta)$ and closes the upper and lower bounds up to logarithmic factors of n, m in this setting.

We provide a proof sketch for Theorem 4.15 here but the full details can be found in Section C.5. It will help to recall the proof structure for Median Elimination (Algorithm 1) which was designed for the (ε, δ)-PAC objective. The proof of the correctness of Median Elimination relies on the fact that after uniformly sampling a set of n arms with a total of n/ε^2 samples, the top half of arms (sorted by empirical means in descending order) contains an ε-good arm with constant probability. By repeatedly invoking this principle with a geometric schedule for the ε_ℓ considered in each round the suboptimality of final surviving arm can be bounded. In Sequential Halving there is no input ε. Instead in each phase $\ell \in [\log_2 n]$ uniform sampling is performed with $t_\ell = \frac{2^\ell T}{n \log_2 n}$ total samples. In analogy with Median Elimination we can consider $\varepsilon_\ell = \sqrt{\frac{n \log_2 n}{2^{\ell-1} T}}$ and arrive at a similar result. This latter result is too weak for our purposes on its own. We can strengthen this result by considering not only what happens to the best arm in each phase, but also what proportion of the top sets of arms carry on to the next

phase.

Theorem 4.15 essentially follows by observing that Sequential Halving keeps at least half of the (m, ε)-good arms from the previous phase for arbitrary $1 \leq m \leq n$. Once again making use of a geometrically shrinking sequence of the ε_ℓ considered in each round, at the start of phase $\ell^* := \log_2 m$ we expect to have at least one of the original (m, ε)-good arms among the remaining n/m. At this point by our previous sketch which bounds the increase in suboptimality across phases of Sequential Halving we expect that after the remaining $(\log_2 n - \log_2 m)$ rounds of sampling are complete we end up with an arm that is no more suboptimal than $\Delta_m + \sqrt{\frac{n \log_2 n}{mT}}$ (if $m \leq n/2$[3]), leading to the expected sample complexity of $\tilde{O}\left(n/m \cdot \varepsilon^{-2}\right)$. It is important to note that m, ε are not inputs to the algorithm, and are purely analysis variables. The flexibility of this result for arbitrary m, ε is the basis for Corollary 4.15.1.

The full proof of Theorem 4.15 can be found in Section C.5 but we show the cornerstone of the proof here, which is essentially a more powerful version of Lemma 3.2. On a given instance let $\text{Top}_m(\varepsilon) := \{i \in [n] : \Delta_i < \Delta_m + \varepsilon\}$ be the set of (m, ε)-good arms, and define $m(\varepsilon) := |\text{Top}_m(\varepsilon)|$. Note that $m(\varepsilon) \geq m$ with equality if and only if $\Delta_{m+1} - \Delta_m > \varepsilon$.

Proposition 4.16 (Uniform Sampling for (m, ε)-optimal arms)**.** *Run uniform exploration with a budget of T and return the $1 \leq s \leq n$ arms with the highest empirical means, $\hat{\mathcal{S}}$, breaking ties arbitrarily. Then for $k \leq s$*

$$\mathbb{P}\left(|\text{Top}_m(\varepsilon) \cap \hat{\mathcal{S}}| \leq k\right) \leq \binom{m}{m-k} \exp\left(-(m-k)\frac{\varepsilon^2 T}{2n}\right) + \binom{n-m(\varepsilon)}{s-k} \exp\left(-(s-k)\frac{\varepsilon^2 T}{2n}\right).$$

The proof of Proposition 4.16 can be found in Section C.3.

Proposition 4.16 tells us that when $m \leq n/2$ and T large enough that the exponential decay dominates the binomial factors, then if $\hat{\mathcal{S}}$ is the top half of arms, the sample complexity of selecting at least $m/2$ of the arms in $\text{Top}_m(\varepsilon)$ with constant probability is of order $n/m \cdot \varepsilon^{-2}$.

We also provide a complementary lower bound for Proposition 4.16 that shows that Proposition 4.16 is essentially tight, and in turn describes the limits of uniform exploration for this setting.

[3]If $m \geq n/2$ then by phase $\ell^* = \log_2 m$ we have at most two arms in the second last phase, with at least one of them being from the original top m arms.

Theorem 4.17 (Lower bound for uniform sampling)**.** *Consider a family $\mathcal{E}(n, m, \varepsilon)$ of all two-level shifted standard normal bandit instances with n arms, of which m arms have mean $\varepsilon > 0$ and all other arms have mean 0. We consider the problem where samples are drawn from some instance $\nu \in \mathcal{E}(n, m, \varepsilon)$ in an off-line, uniform fashion with B samples from each arm (i.e. with a total sampling budget of T samples, $B = T/n$) and a subset of arms $\hat{\mathcal{S}} \subseteq [n]$ of size s is then chosen based only on the observed samples (i.e. without knowledge of ν) in a symmetric fashion. That is to say that the choice of $\hat{\mathcal{S}}$ is independent of the specific indices of the arms chosen.*

Suppose $m \leq s \wedge n/3$ and $s < \frac{6}{11}n$. Then, there exists $\nu \in \mathcal{E}(n, m, \varepsilon)$ such that

$$\mathbb{P}_\nu\left(\left|\hat{\mathcal{S}} \cap \text{Top}_m\right| < k\right) \geq \min\left\{\frac{1}{2},\ 2\left(\frac{6}{5} \cdot \frac{n-s}{s}\right)^{\frac{3}{16}m} \exp\left[-\left(1 + 16 \cdot \frac{\log(en/m)}{\log\left(\frac{6}{5} \cdot \frac{n-s}{s}\right)}\right) mB\varepsilon^2\right]\right\}$$

where $\mathbb{P}_\nu$ is the probability measure induced by performing uniform sampling on instance ν.

The proof of Theorem 4.17 can be found in Section C.4.

While Corollary 4.15.1 shows that Sequential Halving achieves the optimal sample complexity on two-level instances up to log factors, instance-dependent lower bounds for the simple regret (both in in expectation and with high probability) are an open problem. Theorem 4.14 provides some insight as to the correct sample complexity, however it has two drawbacks. One is that it does not show the increase in sample complexity as the desired probability of mis-identification decreases below $1/16$ and so does not clarify what terms in the sample complexity scale with $\log(1/\delta)$. Furthermore, $\mathcal{H}_{\text{low}}(\varepsilon) = \frac{1}{64}\left(-\Delta_{m+1}^{-2} + \frac{1}{m}\sum_{i=m+1}^{n}\Delta_i^{-2}\right)$ is almost certainly looser than the true complexity measure in the sense that the factor of $\frac{1}{m}$ suggests that each of the m ε-good arms on the instance provide an equal reduction in the sample complexity. In reality, even though an instance may have m ε-good arms, if all but one of those m arms have a suboptimality gap only infinitesimally smaller than ε then we expect (e.g. from Theorem 4.2) that the sample complexity lower bound scales like $\Omega\left((n-m)\cdot\varepsilon^{-2}\right)$, not the weaker bound of $\Omega\left((n-2m)/m\cdot\varepsilon^{-2}\right)$ that $\mathcal{H}_{\text{low}}(\varepsilon)$ indicates. We note that on such an instance the bound in Corollary 4.15.1 matches our intuition, as choosing $\tilde{m} > 1$ would necessitate a prohibitively small value in order for $\tilde{\varepsilon}$ to satisfy $\Delta_{\tilde{m}} + \tilde{\varepsilon} \leq \varepsilon$, and so Corollary 4.15.1 provides the expected upper bound of $\tilde{O}\left(n\varepsilon^{-2}\right)$ for the sample complexity of Sequential Halving.

4.4 Anytime Algorithms via the Doubling Trick

A drawback of the algorithms that we have discussed in this chapter is that their design hinges on a known, fixed sampling budget, T. We will refer to this scenario as the known-horizon setting. While there are applications that fit this setting there are also situations where we would like to use an algorithm that can just keep sampling indefinitely while still providing a performance guarantee on its final selection as a function of the number of samples that were seen by the time it was stopped, t. An algorithm that provides such a guarantee is known as an *anytime* algorithm.

Define $\psi_t \coloneqq \psi_{\lfloor t/n \rfloor}$ to be an anytime arm recommendation based on $(\psi_\ell)_{\ell \in \mathbb{N}}$ which is well defined for $t \geq n \log_2 n$.

1: **for** phase $\ell \in \mathbb{N}$ **do**
2: Run Sequential Halving($= (\pi, \psi)$) on $[n]$ with a budget of $T_\ell = 2^\ell \cdot n \log_2 n$ and set ψ_ℓ to the output, ψ_{T_ℓ}
3: **end for**

Algorithm 9: Anytime Sequential Halving

The doubling trick (see e.g. Besson and Kaufmann (2018)) is a well known tool for building anytime online algorithms from their known-horizon counterpart. The idea of the doubling trick is to construct an anytime algorithm by considering a sequence of geometrically growing sampling budgets. This idea is best explained with an example. Anytime Sequential Halving (Algorithm 9) is a simple modification of Sequential Halving (Algorithm 8).

Theorem 4.18. *The output of Algorithm 9 on a 1-subgaussian instance ν after $t \geq n \log_2 n$ rounds of sampling satisfies*

$$\mathbb{P}_{\nu,\pi}\left(\psi_t \neq 1\right) \leq 3 \log_2 n \cdot \exp\left(-\frac{t}{32 \mathcal{H}_2 \log_2 n}\right).$$

Proof. By our conditions on t there must exist some $L \in \mathbb{N}$ such that $2^L \leq t/(n \log_2 n) \leq 2^{L+1}$. It follows that the output ψ_t of Algorithm 9 is the output of a phase no earlier than $\ell = L - 1$, which follows from noting that the total number of samples observed in the first $L-1$ rounds of sampling is $\sum_{\ell=1}^{L-1} 2^\ell \cdot n \log_2 n \leq 2^L \cdot n \log_2 n$. Thus ψ_t corresponds to the output of Sequential Halving (Algorithm 8) in a phase with at least $2^{L-1} \cdot n \log_2 n \geq t/4$ samples by the definition of L, and so the proof follows directly from Theorem 4.4. □

This example demonstrates the ingredients for applying the doubling trick, and shows that constructing order optimal anytime algorithms can be done with only a constant factor increase in sample complexity. That said, algorithms that rely on a doubling trick waste a large portion of the data collected resulting in potentially large constant factors which can lead to poor performance in practical settings. Currently the algorithms with the best guarantees for the fixed budget setting are elimination-style algorithms which are not amenable to the anytime setting without a doubling trick. A nearly-optimal, non-elimination style algorithm that handles the anytime problem setting and making use of all observed data would be an immense practical advancement in pure exploration.

Chapter 5

Conclusion

5.1 Summary

The goal of this thesis was to explore various ways of framing the problem of adaptive data collection in order to investigate the theoretical and practical implications of adopting each framework. In the course of this work we provided novel contributions to the theoretical understanding of several problem settings. Throughout this thesis we have explored the idea that there is no single 'correct' goal for pure exploration. The goal of research into pure exploration theory is not to prescribe problem settings to experimenters, but rather to provide guidance as to what guarantees are or are not possible under different sets of goals and assumptions, some of which may overlap with one another.

Suppose an experimenter has reason to believe that m of the n options they are considering are all admissible choices. If the experimenter were to design a sampling budget with some minimal detectable effect ε in mind, then if they were to frame their task as a best-arm identification problem then they expect to need to run an algorithm with $O\left(n/\varepsilon^{-2}\right)$ samples to ensure a given probability of making a correct selection. On the other hand if they were to make use of the ε-good arm identification framework then they would be able to leverage their knowledge of the number of acceptable arms to cut down their sample size to $\tilde{O}\left((n-m)/m \cdot \varepsilon^{-2}\right)$. Conversely, if the objective of the experimenter is to ensure that they find the best option at some fixed level of confidence then the experimenter can be assured that conducting their experiment with an algorithm such as lil'UCB will ensure that their experiment will terminate nearly as soon as possible while meeting their requirement, however without knowing the smallest suboptimality gap among their options they cannot say how long their

experiment needs to be run for.

Our coverage of pure exploration problems was divided across problems with fixed confidence and fixed budget constraints. In both settings our approach was to first examine the baseline performance of uniform data collection procedures. This first step revealed situations where adaptive data collection could provide improved performance, as well as providing guidance on what mechanisms would lead to this improvement. In the fixed confidence setting we began with the (ε, δ)-PAC problem formulation and showed how to make use of these input parameters to design sample budgets as well as simple mechanisms for introducing adaptivity such as elimination algorithms. Moving to the best arm identification (BAI) setting we were forced to introduce more adaptive algorithms that can distinguish between the top two arms on an instance without prior knowledge of this smallest gap. A key component of the solution to this problem was the development of anytime confidence intervals which allow algorithms to make decisions on the basis of comparisons of sample means at any time while still retaining probabilistic guarantees on their output. We also looked at some limitations of analyzing adaptive algorithms within the (ε, δ)-PAC framework. We discussed the reasons why (ε, δ)-PAC sample complexities seem to contradict our intuitions for finding a good arm among many and we covered alternative problem settings and algorithms that attempt to address these limitations. Finally, leveraging tools built to analyze the statistical complexity of identifying best arms in multi-armed bandits we provided a novel characterization of the limits of adaptation on the related online stochastic convex optimization problem which complement existing results in the area.

Moving on to problems with a fixed budget we revisited the best arm identification setting. While the algorithms that perform well in this setting are relatively simple to analyze, we saw that there is a nuanced relationship between the sample complexity results under various assumptions in this setting. For instance, while uniform sampling attains a nearly optimal error rate on worst-case instances with uniform gaps there is a limit to the performance of adaptive algorithms in this setting; while adaptive algorithms can achieve nearly optimal rates on 'easier' instances where there are a wide range of suboptimality gaps, they pay an extra $\log n$ factor cost in their sample complexity on uniform gap instances. This logarithmic factor cost of adaptation is not present in the corresponding fixed confidence setting. Next we considered the simple regret, describing tight upper and lower bounds for the worst-case setting and followed

this up by taking a close look at the relationship between minimizing the simple regret and minimizing errors in identifying the best arm on instances. The culmination of this section was a novel asymptotic characterization of an explicit trade-off between optimizing the BAI-error and simple regret objectives for oracle-allocation strategies. Finally we introduced the fixed budget ε-good arm identification problem, discussing its relationship to related fixed confidence settings and providing new performance guarantees for an Sequential Halving which show that it achieves the optimal sample complexity up to log factors for certain gap structures.

5.2 Future Directions

5.2.1 Open Problems in Pure Exploration

At this point the majority of pure exploration problems on unstructured bandits have been characterized up to nearly matching instance-dependent upper and lower bounds. The tightly-coupled (ε, δ)-unverifiable sample complexity and fixed budget ε-good arm identification sample complexity remain outliers. Both settings are currently lacking tight upper and lower instance-dependent sample complexity bounds. Such bounds for the fixed budget setting are being actively pursued at the time of this writing. A second area of interest is the development of anytime, fixed budget pure exploration algorithms that operate in a fully-online fashion without a staged elimination structure. The observation that nearly-optimal elimination based algorithms lacked strong empirical performance was a main motivation for the development of lil'UCB by Jamieson et al. (2014). UCB-style algorithms for fixed budget problems were proposed by Audibert, Sebastien Bubeck, and Remi Munos (2010) but either required the values of $\mathcal{H}$ and T in advance, or came without guarantees on the output of the algorithm. Currently, the BUCB algorithm (Katz-Samuels and Jamieson, 2020) comes closest to meeting this design paradigm, however the algorithm requires δ as input, does not come with a performance guarantee for a fixed horizon $T \in \mathbb{N}$, and the existing analysis is believed to be loose on certain gap structures. The design of a fully-online, anytime ε-good arm identification algorithm with a nearly-optimal sample complexity guarantee for the fixed budget setting would likely represent the most practically significant contribution to pure exploration to date.

5.2.2 Inference on Adaptively Collected Data

One should keep in mind that pure exploration is only concerned with a particular adaptive data collection task: collecting samples in order to make an arm selection at the end of a sampling phase. If we are also interested in using the data for other purposes, for example estimating the means of arms, then the use of adaptive data collection algorithms such as bandit algorithms has some drawbacks. For example, samples drawn by these methods can lead to biased estimates of the means of suboptimal arms (Nie et al., 2018), among other related phenomena (Zhang et al., 2020) that make constructing efficient estimators and tight confidence intervals for linear functionals of arm means from such data challenging. A second outcome of adaptive data collection is that (by design) relatively few samples are gathered for certain arms. A consequence of this is that estimates of linear functionals involving these arms are necessarily of higher variance than for estimators based on samples collected from arms chosen uniformly at random. The only way to address this issue of variance is to ensure that a minimum number (or proportion) of samples are drawn from all arms, which in practice amounts to deploying mixture-algorithms that deviate from their adaptive policy to perform uniform sampling on a specified proportion of data collection rounds (Erraqabi et al., 2017). Noting that in the pure exploration setting uniform sampling is associated with lower quality decisions on certain gap structures, and in cumulative regret scenarios entails linear regret, we see that there is a compromise between optimizing the online learning objective and minimizing estimation errors on the collected data.

Intuitively, if statisticians were able to perform a de-biasing procedure on adaptively collected data, then making use of these extra samples would decrease the proportion of uniform sampling required to ensure a prescribed upper bound on estimation errors. Improving said de-biasing procedure would allow for an even larger proportion of rounds to be adaptively sampled while maintaining inference guarantees, in turn diminishing the compromise between the online learning and estimation error objectives. These considerations have motivated a new line of research concerned with developing de-biasing procedures that can recover tight confidence regions for linear functionals of arm means from adaptively sampled data (Khamaru et al., 2021). Current results in this area are asymptotic in nature, and proving a finite-sample characterization of optimal de-biasing procedures in this setting is an open problem.

References

Abe, Naoki and Philip M. Long (1999). "Associative Reinforcement Learning Using Linear Probabilistic Concepts". In: *Proceedings of Machine Learning Research*. ICML. 5

Audibert, Jean-Yves and Sebastien Bubeck (2009). "Minimax Policies for Adversarial and Stochastic Bandits". In: *Proceedings of Machine Learning Research*. COLT. 57

Audibert, Jean-Yves, Sebastien Bubeck, and Remi Munos (2010). "Best Arm Identification in Multi-Armed Bandits". In: *Proceedings of Machine Learning Research*. COLT. 48, 51, 54, 59, 79, 9

Auer, Peter, Nicolò Cesa-Bianchi, and Paul Fischer (2002). "Finite-Time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning*. 12

Aziz, Maryam, Jesse Anderton, Emilie Kaufmann, and Javed Aslam (2018). "Pure Exploration in Infinitely-Armed Bandit Models with Fixed-Confidence". In: *Proceedings of Machine Learning Research*. ALT. 6

Balcan, Maria-Florina, Steve Hanneke, and Jennifer Wortman Vaughan (2010). "The True Sample Complexity of Active Learning". In: *Machine Learning*. 39

Bechhofer, Robert E. (1954). "A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with Known Variances". In: *The Annals of Mathematical Statistics*. 3

Bechhofer, Robert E., Charles W. Dunnett, and Milton Sobel (1954). "A Two-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with a Common Unknown Variance". In: *Biometrika*. 4

Bechhofer, Robert E., J. Kiefer, and Milton Sobel (1968). *Sequential Identification and Ranking Procedures: With Special Reference to Koopman-Darmois Populations*. University of Chicago Press. 4

Besson, Lilian and Emilie Kaufmann (2018). *What Doubling Tricks Can and Can't Do for Multi-Armed Bandits*. arXiv: 1803.06971. 75

Bogachev, V. I. (2007). *Measure Theory*. 1st ed. Berlin ; New York: Springer. 85, 86

Bretagnolle, J. and C. Huber (1978). "Estimation des densités : Risque minimax". In: *Séminaire de Probabilités XII*. 17

Bubeck, Sébastien, Rémi Munos, and Gilles Stoltz (2009). "Pure Exploration in Multi-armed Bandits Problems". In: *Proceedings of Machine Learning Research*. ALT. 12, 55, 56

Bubeck, Sébastien, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári (2011). "X-Armed Bandits". In: *Journal of Machine Learning Research*. 59

Bubeck, Séebastian, Tengyao Wang, and Nitin Viswanathan (2013). "Multiple Identifications in Multi-Armed Bandits". In: *Proceedings of Machine Learning Research*. ICML. 6

Carpentier, Alexandra and Michal Valko (2015). "Simple Regret for Infinitely Many Armed Bandits". In: *Proceedings of Machine Learning Research*. ICML. 6, 59

Chaudhuri, Arghya Roy and Shivaram Kalyanakrishnan (2017). "PAC Identification of a Bandit Arm Relative to a Reward Quantile". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI. 7, 37–39

Chen, Lijie, Jian Li, and Mingda Qiao (2017). "Nearly Instance Optimal Sample Complexity Bounds for Top-k Arm Selection". In: *Proceedings of the Conference on Artificial Intelligence and Statistics*. AISTATS. 33

Chen, Shouyuan, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen (2014). "Combinatorial Pure Exploration of Multi-Armed Bandits". In: *Advances in Neural Information Processing Systems*. NeurIPS. 6

Degenne, Rémy and Wouter M. Koolen (2019). "Pure Exploration with Multiple Correct Answers". In: *Advances in Neural Information Processing Systems*. NeurIPS. 26

Degenne, Rémy, Wouter M. Koolen, and Pierre Ménard (2019). "Non-Asymptotic Pure Exploration by Solving Games". In: *Advances in Neural Information Processing Systems*. 36

Degenne, Rémy, Pierre Menard, Xuedong Shang, and Michal Valko (2020). "Gamification of Pure Exploration for Linear Bandits". In: *Proceedings of Machine Learning Research*. ICML. 6

De Heide, Rianne, James Cheshire, Pierre Ménard, and Alexandra Carpentier (2021). "Bandits with Many Optimal Arms". In: *Advances in Neural Information Processing Systems*. NeurIPS. 6

Erraqabi, Akram, Alessandro Lazaric, Michal Valko, Emma Brunskill, and Yun-En Liu (2017). "Trading off Rewards and Errors in Multi-Armed Bandits". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI. 80

Even-Dar, Eyal, Shie Mannor, and Yishay Mansour (2002). "PAC Bounds for Multi-armed Bandit and Markov Decision Processes". In: *Proceedings of Machine Learning Research*. COLT. 4, 22–24, 50

Garivier, Aurélien and Emilie Kaufmann (2016). "Optimal Best Arm Identification with Fixed Confidence". In: *Proceedings of Machine Learning Research*. COLT. 34, 35, 43, 54, 55

Hazan, Elad, Alexander Rakhlin, and Peter Bartlett (2007). "Adaptive Online Gradient Descent". In: *Advances in Neural Information Processing Systems*. NeurIPS. 42

Jamieson, Kevin, Matthew Malloy, Robert Nowak, and Sébastien Bubeck (2014). "Lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits". In: *Proceedings of Machine Learning Research*. COLT. 28, 29, 32, 33, 79

Jennison, Christopher, Iain M. Johnstone, and Bruce W. Turnbull (1982). "Asymptotically Optimal Procedures for Sequential Adaptive Selection of the Best of Several Normal Means". In: *Statistical Decision Theory and Related Topics III*. 4, 5

Kalvit, Anand and Assaf Zeevi (2021). "A Closer Look at the Worst-case Behavior of Multi-armed Bandit Algorithms". In: *Advances in Neural Information Processing Systems*. NeurIPS. 68

Kalyanakrishnan, Shivaram, Ambuj Tewari, Peter Auer, and Peter Stone (2012). "PAC Subset Selection in Stochastic Multi-Armed Bandits". In: *Proceedings of Machine Learning Research*. ICML. 6

Karnin, Zohar, Tomer Koren, and Oren Somekh (2013). "Almost Optimal Exploration in Multi-Armed Bandits". In: *Proceedings of Machine Learning Research*. ICML. 26, 31, 51, 52, 54

Katz-Samuels, Julian and Kevin Jamieson (2020). "The True Sample Complexity of Identifying Good Arms". In: *Proceedings of the Conference on Artificial Intelligence and Statistics*. AISTATS. 37, 39–41, 79

Kaufmann, Emilie, Olivier Cappe, and Aurelien Garivier (2016). "On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models". In: *The Journal of Machine Learning Research*. 25, 26, 28, 37, 89

Khamaru, Koulik, Yash Deshpande, Lester Mackey, and Martin J. Wainwright (2021). *Near-Optimal Inference in Adaptive Linear Regression*. arXiv: 2107.02266. 80

Kocsis, Levente and Csaba Szepesvári (2006). "Bandit Based Monte-Carlo Planning". In: *Proceedings of Machine Learning Research*. COLT. 3

Lattimore, Tor and Csaba Szepesvári (2020). *Bandit Algorithms*. 1st ed. Cambridge University Press. 6, 12, 17, 34, 35

Mannor, Shie and John N. Tsitsiklis (2003). "Lower Bounds on the Sample Complexity of Exploration in the Multi-armed Bandit Problem". In: *Learning Theory and Kernel Machines*. Springer Berlin Heidelberg. 22, 23, 25

— (2004). "The Sample Complexity of Exploration in the Multi-Armed Bandit Problem". In: *The Journal of Machine Learning Research*. 33

Milton, Abramowitz (1974). *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. 9th ed. New York, NY: Dover Publications. 65

Moulines, Eric and Francis Bach (2011). "Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning". In: *Advances in Neural Information Processing Systems*. NeurIPS. 42, 46

Nemirovsky, A. S. and D. B. Yudin (1983). *Problem Complexity and Method Efficiency in Optimization*. 1st ed. Wiley. 42

Nie, Xinkun, Xiaoying Tian, Jonathan Taylor, and James Zou (Mar. 31, 2018). "Why Adaptively Collected Data Have Negative Bias and How to Correct for It". In: *Proceedings of the Conference on Artificial Intelligence and Statistics*. AISTATS. 80

Paulson, Edward (1952). "An Optimum Solution to the K-Sample Slippage Problem for the Normal Distribution". In: *The Annals of Mathematical Statistics*. 3

— (1964). "A Sequential Procedure for Selecting the Population with the Largest Mean from k Normal Populations". In: *The Annals of Mathematical Statistics*. 4

Ramdas, Aaditya and Aarti Singh (2013). "Algorithmic Connections between Active Learning and Stochastic Convex Optimization". In: *Proceedings of Machine Learning Research*. ALT. 46

Russo, Daniel (2020). "Simple Bayesian Algorithms for Best-Arm Identification". In: *Operations Research*. 68

Tanczos, Ervin, Robert Nowak, and Bob Mankoff (2017). "A KL-LUCB Algorithm for Large-Scale Crowdsourcing". In: *Advances in Neural Information Processing Systems*. NeurIPS. 9

Tsybakov, Alexandre B. (2009). *Introduction to Nonparametric Estimation*. 1st ed. New York ; London: Springer. 18

Wainwright, Martin J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. 1st ed. Cambridge University Press. 13, 18, 62

Zhang, Kelly W, Lucas Janson, and Susan A Murphy (2020). "Inference for Batched Bandits". In: *Advances in Neural Information Processing Systems*. NeurIPS. 80

Zhong, Zixin, Wang Chi Cheung, and Vincent Y. F. Tan (2022). *Achieving the Pareto Frontier of Regret Minimization and Best Arm Identification in Multi-Armed Bandits*. arXiv: 2110.08627. 33

Zhu, Yinglun, Julian Katz-Samuels, and Robert Nowak (May 3, 2022). "Near Instance Optimal Model Selection for Pure Exploration Linear Bandits". In: *Proceedings of the Conference on Artificial Intelligence and Statistics*. AISTATS, pp. 6735–6769. 6

Zhu, Yuancheng, Sabyasachi Chatterjee, John C Duchi, and John Lafferty (2016). "Local Minimax Complexity of Stochastic Convex Optimization". In: *Advances in Neural Information Processing Systems*. NeurIPS. 6, 42, 43, 46

Appendix A

Proofs for Chapter 2

A.1 Proof of Lemma 2.6

We begin with some notation. For a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{F}/\mathcal{G}$-measurable random variable X we define the push-forward of $\mathbb{P}$ under X by $\mathbb{P}_X(A) \coloneqq \mathbb{P}\left(X^{-1}(A)\right)$ for $A \in \mathcal{G}$. We write $\mathbb{P}_{|\mathcal{H}}$ to denote the restriction of the measure $\mathbb{P}$ to a sub σ-algebra $\mathcal{H}$. Before we continue to the proof of Lemma 2.6 we restate some results which can be found in most tomes on measure theory e.g. Bogachev (2007).

Lemma A.1. *Let $\mathbb{P}$ probability measure on measurable space $(\Omega, \mathcal{F})$ and let $\mathcal{H} \subseteq \mathcal{F}$ be a sub σ-algebra of $\mathcal{F}$. Then for any $\mathcal{H}$-measurable random variable X we have*

$$\mathbb{E}_{\mathbb{P}_{|\mathcal{H}}}[X] = \mathbb{E}_{\mathbb{P}}[X]. \tag{A.1}$$

More generally, for nested σ-algebras $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$ we have

$$\mathbb{E}_{\mathbb{P}_{|\mathcal{H}}}[X] = \mathbb{E}_{\mathbb{P}_{|\mathcal{G}}}[X]. \tag{A.2}$$

The lemma follows by the definition of the expectation in terms of the Lebesgue integral, the monotone convergence theorem and by definition of the restriction of a measure.

In the same setting as Lemma A.1, recall that the conditional expectation of a $\mathcal{F}$-measurable random variable Y on Ω, $\mathbb{E}_{\mathbb{P}}[Y|\mathcal{H}]$, is any $\mathcal{H}$-measurable random variable on Ω that satisfies

$$\int_H \mathbb{E}_{\mathbb{P}}[Y|\mathcal{H}]\, \mathrm{d}\mathbb{P} = \int_H Y \mathrm{d}\mathbb{P}, \quad \forall H \in \mathcal{H}. \tag{A.3}$$

Lemma A.2. *In the same setting as, Lemma A.1 it follows by a similar argument that*

$$\mathbb{E}_{\mathbb{P}_{|\mathcal{H}}}[Y|\mathcal{H}] = \mathbb{E}_{\mathbb{P}}[Y|\mathcal{H}] \tag{A.4}$$

on $\mathbb{P}$ almost surely.

Suppose that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$, i.e. $\mathbb{P} \ll \mathbb{Q}$. Then by the Radon-Nikodym theorem (Bogachev, 2007, Theorem 3.2.2) there exists a $\mathbb{Q}$-integrable random variable $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} : \Omega \to [0, \infty)$ for which

$$\mathbb{P}(A) = \int_A \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} \mathrm{d}\mathbb{Q}, \quad \forall A \in \mathcal{F}, \tag{A.5}$$

and this random variable is uniquely defined almost surely on $\mathbb{Q}$.

We will make use of the relationship between the Radon-Nikodym derivative of probability measures $\mathbb{P}, \mathbb{Q}$ on $(\Omega, \mathcal{F})$ and the derivatives of their respective restrictions to the sub σ-algebra $\mathcal{H} \subseteq \mathcal{F}$:

Lemma A.3. *For $\mathbb{P}, \mathbb{Q}$ and $\mathcal{H}$ as listed above,*

$$\frac{\mathrm{d}\mathbb{P}_{|\mathcal{H}}}{\mathrm{d}\mathbb{Q}_{|\mathcal{H}}} = \mathbb{E}_{\mathbb{Q}} \left[\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} \,\middle|\, \mathcal{H} \right] \tag{A.6}$$

almost surely on $\mathbb{Q}$. More generally, if we consider nested σ-algebras $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$

$$\frac{\mathrm{d}\mathbb{P}_{|\mathcal{H}}}{\mathrm{d}\mathbb{Q}_{|\mathcal{H}}} = \mathbb{E}_{\mathbb{Q}_{|\mathcal{G}}} \left[\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} \,\middle|\, \mathcal{H} \right] \tag{A.7}$$

almost surely on $\mathbb{Q}$.

The lemma follows from the uniqueness of the Radon-Nikodym along with the definition of conditional expectation.

Finally we note a relationship between integrals on push-forward measures and their counterparts on the original probability space:

Theorem A.4 (Bogachev (2007, Theorem 3.6.1))**.** *Given probability space $(\Omega_1, \mathcal{F}_1, \mathbb{P})$ and a $\mathcal{F}_1/\mathcal{F}_2$-measurable random variable X between $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, then for any $\mathcal{F}_2$-measurable function f it holds that*

$$\mathbb{E}_{\mathbb{P}} \left[f(X) \right] := \int_{\Omega_1} (f \circ X)(\omega) \mathrm{d}\mathbb{P}(\omega) = \int_{\Omega_2} f(x) \mathrm{d}\mathbb{P}_X(x). \tag{A.8}$$

More generally, for $A \subseteq \Omega_1$ it holds that

$$\int_A (f \circ X)(\omega) \mathrm{d}\mathbb{P}(\omega) = \int_{X(A)} f(x) \mathrm{d}\mathbb{P}_X(x). \tag{A.9}$$

Combining this result with the Radon-Nikodym theorem leads us to a corollary:

Corollary A.4.1. *In the setting of Theorem A.4 let $\mathbb{Q}$ be a probability measure on $(\Omega_1, \mathcal{F}_1)$ for which $\mathbb{P} \ll \mathbb{Q}$. Then it holds that*

$$\frac{\mathrm{d}\mathbb{P}_{|\sigma(X)}}{\mathrm{d}\mathbb{Q}_{|\sigma(X)}} = \frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mathbb{Q}_X} \circ X \tag{A.10}$$

on $\mathbb{Q}$ almost surely.

We are now ready to prove our result. We consider probability measures $\mathbb{P}, \mathbb{Q}$ on the measurable space $(\Omega, \mathcal{F})$ of interaction histories over an infinite number of rounds of interaction between a bandit algorithm $\mathsf{Alg} = (\pi, \tau, \psi)$ and n-armed bandit instances bandit instances ν, ν' respectively. We adopt a pre-drawn observation-table model in which $X_{s,i}$ is the value we observe the s^{th} time that we sample from arm $i \in [n]$, i.e. $X_t = X_{N_{A_t}(t+1),t}$. We assume that ν_i, ν_i' are mutually absolutely continuous for all $i \in [n]$, and that here τ is a stopping time with respect to the filtration $(\mathcal{F}_t)_t \geq 1$, with $\mathcal{F}_t \coloneqq \sigma(A_1, X_1, \ldots, A_t, X_t)$ for which $\mathbb{E}_{\mathbb{P}}[\tau] < +\infty$. We can now define the stopped σ-algebra $\mathcal{F}_\tau = \{A \in \mathcal{F} : A \cap \{\tau \leq t\} \in \mathcal{F}_t, \forall t \geq 1\}$, which is a sub σ-algebra of $\mathcal{F}$. We want to show that for any $\mathcal{F}_\tau/\mathcal{G}$-measurable map X between $(\Omega, \mathcal{F}_\tau)$ and $(\mathcal{X}, \mathcal{G})$

$$\mathrm{KL}(\mathbb{P}_X, \mathbb{Q}_X) \leq \sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}}[N_i(\tau)] \, \mathrm{KL}(\nu_i, \nu_i'). \tag{A.11}$$

We split the proof into three components. First we show that the KL-divergence between $\mathbb{P}_X$ and $\mathbb{Q}_X$ is equal to the KL-divergence between $\mathbb{P}_{|\sigma(X)}$ and $\mathbb{Q}_{|\sigma(X)}$, where $\sigma(X) \coloneqq \{X^{-1}(A) : A \in \mathcal{G}\}$ is the σ-algebra generated by X, which is the smallest σ-algebra for which X is measurable. We then derive a data-processing inequality to show that the KL-divergence between $\mathbb{P}_{|\sigma(X)}$ and $\mathbb{Q}_{|\sigma(X)}$ is smaller than that between $\mathbb{P}_{|\mathcal{F}_\tau}$ and $\mathbb{Q}_{|\mathcal{F}_\tau}$. We conclude by identifying the Radon-Nikodym derivative $\frac{\mathrm{d}\mathbb{P}_{|\mathcal{F}_\tau}}{\mathrm{d}\mathbb{Q}_{|\mathcal{F}_\tau}}$ in terms of the Radon-Nikodym derivatives of $\frac{\mathrm{d}\mathbb{P}_{|\mathcal{F}_t}}{\mathrm{d}\mathbb{Q}_{|\mathcal{F}_t}}$ for $t \geq 1$ and isolating the expectation of $N_i(\tau)$ for each arm $i \in [n]$.

Proof. As promised we begin with a lemma equating KL-divergences between push-forward measures and restrictions measures to σ-algebras generated by a random variable:

Lemma A.5.

$$\mathrm{KL}(\mathbb{P}_X, \mathbb{Q}_X) = \mathrm{KL}(\mathbb{P}_{|\sigma(X)}, \mathbb{Q}_{|\sigma(X)}). \tag{A.12}$$

Proof. Assume that $\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X} < +\infty$. Then we have

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_X, \mathbb{Q}_X\right) &= \int_{\mathcal{X}} -\log\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right) \mathrm{d}\mathbb{P}_X \\
&= \mathbb{E}_{\mathbb{P}}\left[-\log\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right)\right] && \text{(Theorem A.4)} \\
&= \mathbb{E}_{\mathbb{P}_{|\sigma(X)}}\left[-\log\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}\right)\right] \\
&\quad \text{(noting that } \tfrac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X} \text{ is } \sigma(X)\text{-measurable and using Lemma A.1)} \\
&= \int_{\Omega} -\log\left(\frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X} \circ X\right) \mathrm{d}\mathbb{P}_{|\sigma(X)} \\
&= \int_{\Omega} -\log\left(\frac{\mathrm{d}\mathbb{Q}_{|\sigma(X)}}{\mathrm{d}\mathbb{P}_{|\sigma(X)}}\right) \mathrm{d}\mathbb{P}_{|\sigma(X)} && \text{(Corollary A.4.1)} \\
&= \mathrm{KL}(\mathbb{P}_{|\sigma(X)}, \mathbb{Q}_{|\sigma(X)}).
\end{aligned}$$

□

We can then immediately chain Lemma A.5 with a data-processing inequality:

Lemma A.6.

$$\mathrm{KL}(\mathbb{P}_{|\sigma(X)}, \mathbb{Q}_{|\sigma(X)}) \leq \mathrm{KL}(\mathbb{P}_{|\mathcal{F}_\tau}, \mathbb{Q}_{|\mathcal{F}_\tau}). \tag{A.13}$$

Proof. The proof follows by noting that $\sigma(X)$ is a sub σ-algebra of $\mathcal{F}_\tau$, which is a direct result the assumption that X is $\mathcal{F}_\tau$-measurable. We prove the inequality from right to left:

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{|\mathcal{F}_\tau}, \mathbb{Q}_{|\mathcal{F}_\tau}) &= \mathbb{E}_{\mathbb{P}_{|\mathcal{F}_\tau}}\left[\mathbb{E}_{\mathbb{P}_{|\mathcal{F}_\tau}}\left[-\log\left(\frac{\mathrm{d}\mathbb{Q}_{|\mathcal{F}_\tau}}{\mathrm{d}\mathbb{P}_{|\mathcal{F}_\tau}}\right) \middle| \sigma(X)\right]\right] \\
&\quad (\sigma(X) \subseteq \mathcal{F}_\tau \text{ and the tower rule for expectations}) \\
&= \mathbb{E}_{\mathbb{P}_{|\sigma(X)}}\left[\mathbb{E}_{\mathbb{P}_{|\mathcal{F}_\tau}}\left[-\log\left(\frac{\mathrm{d}\mathbb{Q}_{|\mathcal{F}_\tau}}{\mathrm{d}\mathbb{P}_{|\mathcal{F}_\tau}} \middle| \sigma(X)\right)\right]\right] \\
&\quad \text{(noting that the conditional expectation is } \sigma(X)\text{-measurable and using Lemma A.2)} \\
&\geq \mathbb{E}_{\mathbb{P}_{|\sigma(X)}}\left[-\log\left(\mathbb{E}_{\mathbb{P}_{|\mathcal{F}_\tau}}\left[\left(\frac{\mathrm{d}\mathbb{Q}_{|\mathcal{F}_\tau}}{\mathrm{d}\mathbb{P}_{|\mathcal{F}_\tau}} \middle| \sigma(X)\right)\right]\right)\right] && \text{(Jensen's inequality)} \\
&= \mathbb{E}_{\mathbb{P}_{|\sigma(X)}}\left[-\log\left(\frac{\mathrm{d}\mathbb{Q}_{|\sigma(X)}}{\mathrm{d}\mathbb{P}_{|\sigma(X)}}\right)\right] && \text{(Lemma A.3)} \\
&= \mathrm{KL}(\mathbb{P}_{|\sigma(X)}, \mathbb{Q}_{|\sigma(X)}). && \text{(since we assume that this last term is finite)}
\end{aligned}$$

□

Next we write

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{|\mathcal{F}_\tau}, \mathbb{Q}_{|\mathcal{F}_\tau}) &= \mathbb{E}_{\mathbb{P}_{|\mathcal{F}_\tau}}\left[-\log\left(\frac{\mathrm{d}\mathbb{Q}_{|\mathcal{F}_\tau}}{\mathrm{d}\mathbb{P}_{|\mathcal{F}_\tau}}\right)\right] \\
&= \mathbb{E}_{\mathbb{P}_{|\mathcal{F}_\tau}}\left[-\log\left(\mathbb{E}_{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\,\middle|\,\mathcal{F}_\tau\right]\right)\right] && \text{(Lemma A.3)} \\
&= \mathbb{E}_{\mathbb{P}}\left[-\log\left(\mathbb{E}_{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\,\middle|\,\mathcal{F}_\tau\right]\right)\right]. && \text{(Lemma A.2)}
\end{aligned}$$

The penultimate step in the proof is to construct $-\log\left(\mathbb{E}_{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\,\middle|\,\mathcal{F}_\tau\right]\right)$. We will make use of a result by Kaufmann et al. (2016). Define the $\mathcal{F}_t$-measurable random variable

$$L_t = L_t(A_1, X_1, \ldots, A_t, X_t) \coloneqq \sum_{i=1}^{n}\sum_{s=1}^{t} \mathbb{1}\{A_s = i\}\log\left(\frac{p_i(X_s)}{p_i'(X_s)}\right) \tag{A.14}$$

for $t \geq 1$, where $p_i(x), p_i'(x)$ are the respective densities of ν_i, ν_i' with respect to $\lambda = \sum_{i=1}^{n}\nu_i + \nu_i'$, where these densities are well defined by our assumption that the ν_i, ν_i' are mutually absolutely continuous. With this definition L_τ is an $\mathcal{F}_\tau$-measurable random variable. Recall that by the definition of conditional expectation and the Radon-Nikodym derivative, $-\log\left(\mathbb{E}_{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\,\middle|\,\mathcal{F}_\tau\right]\right)$ is any $\mathcal{F}_\tau$-measurable random variable $f : \Omega \to (-\infty, +\infty)$ with the property that

$$\mathbb{Q}(A) = \int_A \exp(-f)\mathrm{d}\mathbb{P} \tag{A.15}$$

for every $A \in \mathcal{F}_\tau$, and where this random variable is uniquely defined on $\mathbb{P}$ almost surely. The next lemma shows that L_τ is the random variable that we are looking for.

Lemma A.7 ((Kaufmann et al., 2016, Lemma 18))**.** *For every event $A \in \mathcal{F}_\tau$,*

$$\mathbb{Q}(A) = \int_A \exp(-L_\tau)\,\mathrm{d}\mathbb{P}. \tag{A.16}$$

Putting everything together, we have

$$
\begin{aligned}
\mathrm{KL}(\mathbb{P}_X, \mathbb{Q}_X) &= \mathrm{KL}(\mathbb{P}_{|\sigma(X)}, \mathbb{Q}_{|\sigma(X)}) \\
&\leq \mathrm{KL}(\mathbb{P}_{|\mathcal{F}_\tau}, \mathbb{Q}_{|\mathcal{F}_\tau}) \\
&= \mathbb{E}_{\mathbb{P}}\left[-\log\left(\mathbb{E}_{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \,\middle|\, \mathcal{F}_\tau\right]\right)\right] \\
&= \mathbb{E}_{\mathbb{P}}\left[L_\tau\right] \\
&= \mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^{n}\sum_{t=1}^{\tau} \mathbb{1}\{A_t = i\} \log\left(\frac{p_i(X_t)}{p'_i(X_t)}\right)\right] && \text{(definition of } L_\tau\text{)} \\
&= \mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^{n}\sum_{t=1}^{N_i(\tau)} \log\left(\frac{p_i(X_{t,i})}{p'_i(X_{t,i})}\right)\right] \\
&\qquad \text{(collecting terms and using the observation-table probability model)} \\
&= \sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{N_i(\tau)} \log\left(\frac{p_i(X_{t,i})}{p'_i(X_{t,i})}\right)\right] && \text{(linearity of expectation)} \\
&= \sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{\infty} \mathbb{1}\{N_i(\tau) \geq t\} \log\left(\frac{p_i(X_{t,i})}{p'_i(X_{t,i})}\right)\right] \\
&= \sum_{i=1}^{n}\sum_{t=1}^{\infty} \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}\{N_i(\tau) \geq t\} \log\left(\frac{p_i(X_{t,i})}{p'_i(X_{t,i})}\right)\right] \\
&\qquad \text{(assumption of finite } \mathrm{KL}(\nu_i, \nu'_i) \text{ and dominated convergence)} \\
&= \sum_{i=1}^{n}\sum_{t=1}^{\infty} \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{P}}\left[\mathbb{1}\{N_i(\tau) \geq t\} \log\left(\frac{p_i(X_{t,i})}{p'_i(X_{t,i})}\right) \,\middle|\, \mathcal{F}_{t-1}\right]\right] && \text{(tower rule)} \\
&= \sum_{i=1}^{n}\sum_{t=1}^{\infty} \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}\{N_i(\tau) \geq t\} \cdot \mathbb{E}_{\mathbb{P}}\left[\log\left(\frac{p_i(X_{t,i})}{p'_i(X_{t,i})}\right) \,\middle|\, \mathcal{F}_{t-1}\right]\right] \\
&\text{(independence of } X_{t,i} \text{ and } \mathcal{F}_{t-1}\text{, noting that } \mathbb{1}\{N_i(\tau) \geq t\} \,(= 1 - \mathbb{1}\{N_i(\tau) \leq t-1\}) \text{ is } \mathcal{F}_{t-1}\text{-measurable.)} \\
&= \sum_{i=1}^{n}\sum_{t=1}^{\infty} \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}\{N_i(\tau) \geq t\} \cdot \mathrm{KL}(\nu_i, \nu'_i)\right] \\
&\qquad \text{(i.i.d. sampling of obervations from arm } i \in [n]\text{)} \\
&= \sum_{i=1}^{n}\sum_{t=1}^{\infty} \mathbb{P}\left(N_i(\tau) \geq t\right) \mathrm{KL}(\nu_i, \nu'_i) && \text{(identity)} \\
&= \sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}}\left[N_i(\tau)\right] \mathrm{KL}(\nu_i, \nu'_i) \\
&\qquad \text{(} \mathbb{E}_{\mathbb{P}}\left[N_i(\tau)\right] \leq \mathbb{E}_{\mathbb{P}}\left[\tau\right] < +\infty \text{ and using the tail sum formula)}
\end{aligned}
$$

It can be verified that if the right-hand side is infinite then so is the left-hand side. □

Appendix B

Proofs for Chapter 3

B.1 Proof of Theorem 3.16

The proof follows the same general form as Theorem 3.10 in the multi-armed bandit setting.

Proof. Let $g \in \mathcal{E}_{\text{alt}}(f, \varepsilon)$. By the definition of (ε, δ)-PAC algorithms we have

$$\begin{aligned} 2\delta &\geq \mathbb{P}_{f,\pi}(\Delta_f(\psi_\tau) > \varepsilon) + \mathbb{P}_{g,\pi}(\Delta_g(\psi_\tau) > \varepsilon) \\ &\geq \mathbb{P}_{f,\pi}(\Delta_g(\psi_\tau) \leq \varepsilon) + \mathbb{P}_{g,\pi}(\Delta_g(\psi_\tau) > \varepsilon). && \text{(by our choice of } g \in \mathcal{E}^n_{\text{alt}}(f,\varepsilon)) \end{aligned}$$

Now we can apply the Bretagnolle-Huber inequality and the chain rule for Kullback-Leibler (KL) divergences with stopping times to get

$$\begin{aligned} 2\delta &\geq \mathbb{P}_{f,\pi}(\Delta_g(\psi_\tau) \leq \varepsilon) + \mathbb{P}_{g,\pi}(\Delta_g(\psi_\tau) > \varepsilon) \\ &\geq \frac{1}{2}\exp\left(-\mathbb{E}_{f\pi}\left[\mathrm{KL}\left(\mathbb{P}_{f,\pi,\tau}, \mathbb{P}_{g,\pi,\tau}\right)\right]\right) && \text{(Bretagnolle-Huber inequality)} \\ &\geq \frac{1}{2}\exp\left(-\mathbb{E}_{f\pi}\left[\sum_{t=1}^{\tau} \mathrm{KL}\left(\mathbb{P}_{f,\pi}(X_t), \mathbb{P}_{g,\pi}(X_t)\right)\right]\right) && \text{(chain rule for KL-divergences)} \\ &\geq \frac{1}{2}\exp\left(-\mathbb{E}_{f\pi}\left[\sum_{t=1}^{\tau} \frac{\|\partial f(X_t) - \partial g(X_t)\|_2^2}{2}\right]\right) && \text{(multivariate normal KL)} \\ &\geq \frac{1}{2}\exp\left(-\mathbb{E}_{f\pi}[\tau] \sup_{x\in\mathcal{D}} \frac{\|\partial f(x) - \partial g(x)\|_2^2}{2}\right). \end{aligned}$$

Rearranging the display above and tightening the bound with a specific choice of g,

$$\begin{aligned} \mathbb{E}_{f,\pi}[\tau] &\geq \left(\frac{\inf_{g\in\mathcal{E}^n_{\text{alt}}(f,\varepsilon)} \sup_{x\in\mathcal{D}} \|\partial f(x) - \partial g(x)\|_2^2}{2}\right)^{-1} \log\left(\frac{1}{4\delta}\right) \\ &= c^*_f(\varepsilon) \log\left(\frac{1}{4\delta}\right) \end{aligned}$$

with $c_f^*(\varepsilon)^{-1} = \inf_{g \in \mathcal{E}_{\text{alt}}(f,\varepsilon)} \sup_{x \in \mathcal{D}} \frac{\|\partial f(x) - \partial g(x)\|_2^2}{2}$. $\square$

Appendix C

Proofs for Chapter 4

C.1 Proof of Theorem 4.2

Proof. The proof follows from a specialization and subsequent strengthening (in terms of the n-dependence) of the lower-bound presented by Audibert, Sebastien Bubeck, and Remi Munos (2010) to our instance where all gaps are equal, and makes use of similar notation and propositions (some presented without proof) in the form that is used in the original lower-bound:

It will be useful to consider an empirical estimate of the KL-divergence, obtained from an average of $t \in \mathbb{N}$ samples from arm $i \in [n]$,

$$\widehat{\mathrm{KL}}_{i,t}\left(\mathcal{B}(p), \mathcal{B}(q)\right) \coloneqq \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}\{X_{i,s}=1\}\log\left(\frac{p}{q}\right) + \mathbb{1}\{X_{i,s}=0\}\log\left(\frac{1-p}{1-q}\right), \tag{C.1}$$

where we abuse our earlier notation to let $X_{i,t}$ denote the t^{th} sample observed from arm i. Note that when $X_{i,s} \sim \mathcal{B}(p)$, $\mathbb{E}\left[\widehat{\mathrm{KL}}_{a,t}\left(\mathcal{B}(p), \mathcal{B}(q)\right)\right] = \mathrm{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right)$.

Lemma C.1 (Bandit change of measure)**.** *Let ν and ν' be two bandit instances that differ only in the distribution of arm $j \in [n]$. Then for an arbitrary learner $\mathsf{Alg} = (\pi, \psi)$ and any event $\mathcal{E} \in \mathcal{F}$,*

$$\mathbb{P}_{\nu,\pi}\left(\mathcal{E}\right) = \mathbb{E}_{\nu,\pi}\left[\mathbb{1}\{\mathcal{E}\}\right] = \mathbb{E}_{\nu',\pi}\left[\mathbb{1}\{\mathcal{E}\}\exp\left(-N_a(T)\,\widehat{\mathrm{KL}}_{a,N_j(T)}\left(\nu'_j, \nu_j\right)\right)\right]. \tag{C.2}$$

Lemma C.2 (A maximal form of the Azuma-Hoeffding inequality)**.** *Let $(X_t)_{t=1}^T$ be a sequence of centered and bounded random variables s.t. $|X_t| \leq b$ for some $b > 0$. Then $M_t = \sum_{s=1}^t X_s$ is martingale w.r.t. the filtration on $(X_t)_{t=1}^T$, with bounded differences $|M_t - M_{t-1}| \leq b$, and we have for arbitrary $T \geq 1$, $\varepsilon > 0$,*

$$\mathbb{P}\left(\max_{s\leq n} M_s > \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{Tb^2}\right). \tag{C.3}$$

Recall that the environment class being considered consists of permutations of an instance with Bernoulli arms, with means $\mu_1 = 1/2$, $(\mu_i)_{i=2}^n = p$ for some $p \in (0, 1/2)$. Let $\nu = (\mu_i)_{i=1}^n$, and $\nu' = (\mu_n, \mu_2, \ldots, \mu_n)$, so that ν' is a new bandit instance obtained by changing the distribution of the first arm from $\mathcal{B}(\mu_1)$ to $\mathcal{B}(\mu_n)$.

Reasoning about under-sampled arms: Consider the random variable $Z \in \arg\min_{i \neq \psi_T} N_i(T)$, where in the case of multiple least-played arms Z is drawn uniformly from the subset of arms

$$\{i \neq \psi_T : N_i(T) \leq \min_{j \in [n] \setminus \{\psi_T\}} N_j(T)\}.$$

Z corresponds to a least-played arm that is not recommended by $\mathsf{Alg} = (\pi, \psi)$ in the final round. It follows that $N_Z(T) \leq \frac{1}{n-1} \sum_{i \neq \psi_T} N_i(T)$.

An event on which the empirical KL-divergence concentrates uniformly: Let

$$\begin{aligned} \xi = \Big\{ & \forall t \in [T], i \neq 1, j \in [n], \\ & \widehat{\mathrm{KL}}_{i,t}(\nu_i, \nu_i) \leq \mathrm{KL}(\nu_i, \nu_i) + 2\log(3)\sqrt{\frac{(\log 6n^2)}{T}}, \\ & \widehat{\mathrm{KL}}_{1,t}(\nu_n, \nu_i) \leq \mathrm{KL}(\nu_n, \nu_i) + 2\log(3)\sqrt{\frac{(\log 6n^2)}{T}} \Big\} \end{aligned} \tag{C.4}$$

be an event on which the empirical KL-divergence concentrates uniformly over arms and rounds of sampling. Then for any learner Alg, $\mathbb{P}_{\nu'}[\xi] \geq 2/3$, where the proof follows by noticing that since the Bernoulli mean parameters of all arms lie in an interval bounded away from 0 the empirical KL cannot drift away from the true KL too quickly (which is also bounded given the constraints on the arm parameters).

More formally, we have the following: for all $i \in [n], \mu_i \in [\frac{1}{2}, \frac{3}{4}, i, j \in [n]$

$$\left| \log\left(\frac{\mu_i}{\mu_j}\right) \right| \leq \log(8/3) \leq \log 3, \tag{C.5}$$

and

$$\left| \log\left(\frac{1-\mu_i}{1-\mu_j}\right) \right| \leq \log\left(\frac{1/2}{1/4}\right) \leq \log 3. \tag{C.6}$$

It follows that on ν', for $i \neq 1, j \in [n]$, $M_t \coloneqq t\left[\widehat{\mathrm{KL}}_{i,t}\left(\nu_i, \nu_j\right) - \mathrm{KL}\left(\nu_i, \nu_j\right)\right]$ is a martingale sequence with bounded differences:

$$\begin{aligned}
|M_t - M_{t-1}| &= \left|\mathbb{1}\{X_{i,t} = 1\}\log\left(\frac{\mu_i}{\mu_j}\right) + \mathbb{1}\{X_{i,t} = 0\}\log\left(\frac{1-\mu_i}{1-\mu_j}\right) - \mathrm{KL}\left(\nu_i, \nu_j\right)\right| \\
&\leq \left|\mathbb{1}\{X_{i,t} = 1\}\log\left(\frac{\mu_i}{\mu_j}\right) + \mathbb{1}\{X_{i,t} = 0\}\log\left(\frac{1-\mu_i}{1-\mu_j}\right)\right| + \left|\mathrm{KL}\left(\nu_i, \nu_j\right)\right| \\
&\leq 2\log 3 \qquad \text{(C.7)}
\end{aligned}$$

for $t \in [T]$, where $M_0 \coloneqq 0$.

From Lemma C.2 we have that for any learner Alg, and choice of arms $i \neq 1, j \in [n]$,

$$\mathbb{P}_{\nu',\mathsf{Alg}}\left(\max_{t\in[T]}\left[\widehat{\mathrm{KL}}_{i,t}\left(\nu_i, \nu_j\right) - \mathrm{KL}\left(nu_i, \nu_j\right)\right] > 2\log(3)\sqrt{\frac{\log(6Tn^2)}{T}}\right) \leq \frac{1}{6n^2}. \qquad \text{(C.8)}$$

By the same argument, once again on ν' we have

$$\mathbb{P}_{\nu',\mathsf{Alg}}\left(\max_{t\in[T]}\left[\widehat{\mathrm{KL}}_{1,t}\left(\nu_n, \nu_j\right) - \mathrm{KL}\left(\nu_n, \nu_j\right)\right] > 2\log(3)\sqrt{\frac{\log(6Tn^2)}{T}}\right) \leq \frac{1}{6n^2}, \qquad \text{(C.9)}$$

and so taking a union bound over $i \neq 1, j \in [n]$ we have $\mathbb{P}_{\nu',\mathsf{Alg}}[\xi] \geq 2/3$.

A change of measure argument: By boolean algebra identities we have $\xi = \bigcap_{z\in[n]}\left(\xi \cap \{Z = z\}\right)$ and so $\sum_{z\in[n]}\mathbb{P}_{\nu',\mathsf{Alg}}\left[\xi \cap \{Z = z\}\right] \geq 2/3$, and there must exist some $\tilde{z} \in [n]$ such that $\mathbb{P}_{\nu',\mathsf{Alg}}\left[\xi \cap \{Z = \tilde{z}\}\right] \geq \frac{2}{3(n-1)}$. Let $\mathcal{E} = \xi \cap \{Z = \tilde{z}\}$ for such a $\tilde{z} \in [n]$, and let $\sigma(\nu)$ be a new bandit instance obtained by swapping the distributions of arms 1 and $\tilde{z}$ on instance ν. It will be useful to recall that $\sigma(\nu)$ and ν' differ only on index $\tilde{z}$, and on $\sigma(\nu)$ the distribution at index $\tilde{z}$ is ν_1, whereas all other arms have distribution ν_n. Further, by definition $Z \neq \psi_T$, so $\mathcal{E} = \xi \bigcap \{Z = \tilde{z}\} \subseteq \{\psi_T \neq \tilde{z}\}$. Applying Lemma C.1 we have

$$\begin{aligned}
\mathbb{P}_{\sigma(\nu)}\left(\psi_T \neq \sigma(1)\right) &= \mathbb{E}_{\sigma(\nu),\mathsf{Alg}}\left[\mathbb{1}\{\psi_T \neq \tilde{z}\}\right] \\
&= \mathbb{E}_{\nu',\mathsf{Alg}}\left[\mathbb{1}\{\psi_T \neq \tilde{z}\}\exp\left(-N_{\tilde{z}}(T)\,\widehat{\mathrm{KL}}_{1,N_{\tilde{z}}(T)}\left(\nu_n, \nu_1\right)\right)\right] \\
&\geq \mathbb{E}_{\nu',\mathsf{Alg}}\left[\mathbb{1}\{\mathcal{E}\}\exp\left(-N_{\tilde{z}}(T)\,\widehat{\mathrm{KL}}_{1,N_{\tilde{z}}(T)}\left(\nu_n, \nu_1\right)\right)\right]. \qquad \text{(C.10)}
\end{aligned}$$

Making use of concentration of the empirical KL on ξ, and the upper bound on $N_Z(T)$,

$$\begin{aligned}
\mathbb{P}_{\sigma(\nu)}\left(\psi_T \neq \sigma(1)\right) &\geq \mathbb{E}_{\nu',\mathsf{Alg}}\left[\mathbb{1}\{\mathcal{E}\}\exp\left(-N_Z(T)\mathrm{KL}\left(\nu_n,\nu_1\right) - 2\log(3)\sqrt{N_Z(T)\log 6n^2}\right)\right] \\
&\geq \mathbb{E}_{\nu',\mathsf{Alg}}\left[\mathbb{1}\{\mathcal{E}\}\right]\exp\left(-\frac{T}{(n-1)}\mathrm{KL}\left(\nu_k,\nu_1\right) - 2\log(3)\sqrt{\frac{T}{(n-1)}\log 6n^2}\right) \\
&\geq \frac{2}{3(n-1)}\exp\left(-\frac{T}{(n-1)}\mathrm{KL}\left(\nu_n,\nu_1\right) - 2\log(3)\sqrt{\frac{T}{(n-1)}(\log 6n^2)}\right) \\
&\geq \frac{2}{3(n-1)}\exp\left(-\frac{3\Delta^2 T}{(n-1)} - 2\log(3)\sqrt{\frac{T}{(n-1)}(\log 6n^2)}\right)
\end{aligned}$$

where in the second line the upper bound on $N_Z(T)$ is used twice, and in the last line the KL-divergence term is upper bounded by $3\Delta^2$. □

C.2 Proof of Theorem 4.10

Proof. Let ν be an n-armed shifted standard normal instance with means $(\Delta, 0, 0, \ldots, 0)$ and $n \geq 10$ for some Δ to be specified later, and let $\tilde{\nu}$ be an almost identical instance to ν, with arms 1 and 2 interchanged. Recall that we are considering a symmetric algorithm, (π, ψ), where π corresponds to uniform sampling of arms. We have

$$\begin{aligned}
\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] &= \Delta\sum_{i=2}^{n}\mathbb{P}_{\nu,\pi}\left(\psi_T = i\right) \\
&= (n-1)\Delta\mathbb{P}_{\nu,\pi}\left(\psi_T = 2\right) \\
&\qquad\qquad \text{(identical distributions for arms } i \geq 2 \text{ and equivariance of } \psi_T\text{)} \\
&= (n-1)\Delta\mathbb{E}_{\tilde{\nu},\pi}\left[\mathbb{1}\{\psi_T = 2\}\exp\left(-\sum_{a=1}^{n}\sum_{t=1}^{T/n}\log\left(\frac{\mathrm{d}\mathbb{P}_{\nu,\pi}(X_{a,t}|A_t = a)}{\mathrm{d}\mathbb{P}_{\pi,\tilde{\nu}}(X_{a,t}|A_t = a)}\right)\right)\right]. \\
&\qquad\qquad \text{(change of measure)}
\end{aligned}$$

We consider the event on which the 'empirical KL-divergence' concentrates:

$$\mathcal{E}(\rho) = \left\{\sum_{a=1}^{n}\sum_{t=1}^{T/n}\log\left(\frac{\mathrm{d}\mathbb{P}_{\nu,\pi}(X_{a,t}|A_t = a)}{\mathrm{d}\mathbb{P}_{\pi,\tilde{\nu}}(X_{a,t}|A_t = a)}\right) - (1+\rho)\mathrm{KL}(\nu_a,\tilde{\nu}_a) \leq \frac{1}{\rho}\log(1/\delta)\right\}.$$

Lemma C.3. *For arbitrary $\rho > 0$, $\mathbb{P}_{\tilde{\nu},\pi}(\mathcal{E}(\rho)) \geq 1 - \delta$*

It follows that for fixed $\delta \in (0,1)$, $\rho > 0$

$$\begin{aligned}
\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] &= (n-1)\Delta \mathbb{E}_{\tilde{\nu},\pi}\left[\mathbb{1}\{\psi_T = 2\}\exp\left(-\sum_{a=1}^{n}\sum_{t=1}^{T/n}\log\left(\frac{\mathrm{d}\mathbb{P}_{\nu,\pi}(X_{a,t}|A_t=a)}{\mathrm{d}\mathbb{P}_{\pi,\tilde{\nu}}(X_{a,t}|A_t=a)}\right)\right)\right] \\
&\geq (n-1)\Delta\exp\left(-(1+\rho)\frac{\Delta^2 T}{n} - \frac{1}{\rho}\log(1/\delta)\right)\mathbb{P}_{\tilde{\nu},\pi}\left(\{\psi_T = 2\}\cap\{\mathcal{E}(\rho)\}\right) \\
&\qquad\qquad \text{(uniform exploration and on } \mathcal{E}(\rho)) \\
&\geq (n-1)\Delta\exp\left(-(1+\rho)\frac{\Delta^2 T}{n} - \frac{1}{\rho}\log(1/\delta)\right)\left(\mathbb{P}_{\tilde{\nu},\pi}\left(\psi_T = 2\right) - \delta\right) \\
&= (n-1)\Delta\exp\left(-(1+\rho)\frac{\Delta^2 T}{n} - \frac{1}{\rho}\log(1/\delta)\right)\left(\mathbb{P}_{\nu,\pi}\left(\psi_T = 1\right) - \delta\right). \\
&\qquad\qquad \text{(symmetry)}
\end{aligned}$$

Now consider the case that $\mathbb{P}_{\nu,\pi}\left(\psi_T \neq 1\right) \leq \frac{1}{2}$. Taking $\delta = \frac{1}{4}$ and $\rho = 1$ we have

$$\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] \geq \frac{(n-1)\Delta}{8}\exp\left(-\frac{2\Delta^2 T}{n}\right)\mathbb{P}_{\nu,\pi}\left(\psi_T = 1\right),$$

and finally taking $\Delta = \sqrt{\dfrac{n\log(n-1)}{2n}}$, we have

$$\begin{aligned}
\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] &\geq \frac{1}{8e}\sqrt{\frac{n\log(n-1)}{T}} \\
&\geq \frac{1}{9e}\sqrt{\frac{n\log n}{T}}. \qquad \text{(when } n \geq 10)
\end{aligned}$$

Considering the case when $\mathbb{P}_{\nu,\pi}\left(\psi_T \neq 1\right) > \frac{1}{2}$ and choosing $\Delta = \sqrt{\dfrac{n\log n}{T}}$ we immediately get

$$\mathbb{E}_{\nu,\pi}\left[\Delta_{\psi_T}\right] > \frac{1}{2}\sqrt{\frac{n\log n}{T}},$$

which completes the proof. □

Proof of Lemma C.3. Let

$$\Lambda_t(a) = \left(\frac{\mathrm{d}\mathbb{P}_{\nu,\pi}(X_{a,t}|A_t=a)}{\mathrm{d}\mathbb{P}_{\pi,\tilde{\nu}}(X_{a,t}|A_t=a)}\right) \tag{C.11}$$

and for $\rho > 0$ define

$$H_t = \exp\left(\sum_{a=1}^{n}\rho\left(\sum_{s=1}^{t}\Lambda_s(a) - (1+\rho)\mathrm{KL}(\nu_a,\tilde{\nu}_a)\right)\right), H_0 = 1.$$

The proof follows by showing that $(H_t)_{t\geq 0}$ is a non-negative martingale. For an integrable random variable U, for $t \geq 1$ let

$$\mathbb{E}_{\tilde{\nu},t-1}[U] \coloneqq \mathbb{E}_{\tilde{\nu}}[U|X_{1,1}, X_{2,1}, \ldots, X_{n,1}, X_{2,2}, \ldots, X_{n,t-1}]. \tag{C.12}$$

Then

$$\mathbb{E}_{\tilde{\nu},t-1}[H_t] = H_{t-1} \cdot \mathbb{E}_{\tilde{\nu},t-1}\left[\exp\left(\sum_{a=1}^{n} \rho\Lambda_t(a) - \rho(1+\rho)\mathrm{KL}_{q+\ell}(a)\right)\right].$$

Let $\tilde{\mu}_a, \mu_a$ be the means of arm a on $\tilde{\nu}$ and ν respectively. We have

$$\begin{aligned}
&\mathbb{E}_{\tilde{\nu},t-1}\left[\exp\left(\sum_{a=1}^{n}(\rho\Lambda_t(a) - \rho(1+\rho)\mathrm{KL}_{q+\ell}(a))\right)\right] \\
&= \mathbb{E}_{\tilde{\nu},t-1}\left[\exp\left(\sum_{a=1}^{n}\left(\rho\frac{-2X_{a,t}(\mu'_a-\mu_a)^2 + (\mu'_a)^2 + \mu_a^2}{2} - \rho(1+\rho)\frac{(\mu'_a-\mu_a)^2}{2}\right)\right)\right] \\
&= \mathbb{E}_{\theta',t-1}\left[\exp\left(\sum_{a=1}^{n}\left(\rho\frac{-2\mu'_a(\mu'_a-\mu_a)^2 + (\mu'_a)^2 + \mu_a^2}{2} - \rho(X_{a,t}-\mu'_a)(\mu_a-\mu'_a)\right.\right.\right. \\
&\qquad \left.\left.\left. -\rho(1+\rho)\frac{(\mu'_a-\mu_a)^2}{2}\right)\right)\right] \qquad \text{(refactoring to bring out the } X_{a,t}-\mu'_a \text{ term)} \\
&= \exp\left(\sum_{a=1}^{n}\left(\rho\frac{-2\mu'_a(\mu'_a-\mu_a)^2 + (\mu'_a)^2 + \mu_a^2}{2} + \rho^2\frac{(\mu_a-\mu'_a)^2}{2} - \rho(1+\rho)\frac{(\mu'_a-\mu_a)^2}{2}\right)\right) \\
&\qquad \text{(gaussian moment-generating function)} \\
&= \exp\left(\sum_{a=1}^{n}\left(\rho\frac{(\mu_a-\mu'_a)^2}{2} + \rho^2\frac{(\mu_a-\mu'_a)^2}{2} - \rho(1+\rho)\frac{(\mu'_a-\mu_a)^2}{2}\right)\right) \\
&= 1.
\end{aligned}$$

It follows from Markov's inequality that for $T \geq 0$, $\mathbb{P}_{\tilde{\nu},\pi}(H_T \geq \varepsilon) \leq \frac{1}{\varepsilon}$, and so for $T \geq 0$, $\delta > 0$

$$\mathbb{P}_{\tilde{\nu},\pi}\left(\sum_{a=1}^{n}\sum_{t=1}^{T/n}\Lambda_t(a) - (1+\rho)\mathrm{KL}(\nu_a,\tilde{\nu}_a) \geq \frac{1}{\rho}\log(1/\delta)\right) \leq \delta. \tag{C.13}$$

□

C.3 Proof of Proposition 4.16

Proof. Let $\mathrm{Top}_m(0) \coloneqq \mathrm{Top}_m$. The proof follows from the intuition that in order for no more than k arms from $\mathrm{Top}_m(\varepsilon)$ to be returned in the top s arms, there must have been at least $s-k$ arms from $\overline{\mathrm{Top}}_m(\varepsilon)$ with larger empirical means than the missing arms from $\mathrm{Top}_m(\varepsilon)$. Recall that the sets Top_m and $\overline{\mathrm{Top}}_m(\varepsilon)$ are ε-separated. For a given pair of arms $i \in \mathrm{Top}_m$,

$j \in \overline{\text{Top}}_m(\varepsilon)$ one can verify that $\{\hat{\mu}_i \le \hat{\mu}_j\} \subseteq \{\hat{\mu}_i \le \mu_i - \varepsilon/2\} \cup \{\hat{\mu}_j \ge \mu_j + \varepsilon/2\}$. That is to say that in order for a 'good' arm to have been picked off by a 'bad' arm, either the good arm was underestimated or the bad arm was overestimated. Applying this principle more generally we claim that when

- $|\{i \in \text{Top}_m : \hat{\mu}_i > \mu_i - \varepsilon/2\}| \ge k+1$ and
- $\left|\left\{i \in \overline{\text{Top}}_m(\varepsilon) : \hat{\mu}_i < \mu_i + \varepsilon/2\right\}\right| \ge m(\varepsilon) - s + k + 1$

both hold then $|\text{Top}_m(\varepsilon) \cap \hat{\mathcal{S}}| > k$. One way to see this is to realize that when the first event holds then there are at least $k+1$ 'robust' good arms, i.e. arms that cannot be picked off by slightly over-estimated bad arms. Combined with the second event that limits the number of over-estimated bad arms to no more than $s-k-1$, we see that these $k+1$ robust good arms must be included in $\hat{\mathcal{S}}$, since there are not enough overestimated bad arms to displace them.

The proof follows by a reverse union bound on the probability of either of these events failing. Considering each term individually:

$$\begin{aligned}
&\mathbb{P}\left(|\{i \in \text{Top}_m : \hat{\mu}_i > \mu_i - \varepsilon/2\}| < k+1\right) \\
&= \mathbb{P}\left(|\{i \in \text{Top}_m : \hat{\mu}_i \le \mu_i - \varepsilon/2\}| \ge m-k\right) \\
&= \mathbb{P}\left(\exists \mathcal{A} \subset \text{Top}_m : |\mathcal{A}| = m-k, \text{ and } \forall i \in \mathcal{A}, \hat{\mu}_i \le \mu_i - \varepsilon/2\right) \\
&\le \binom{m}{m-k} \exp\left(-(m-k)\frac{\varepsilon^2 T}{2n}\right) \qquad \text{(Hoeffding and uniform sampling)}
\end{aligned}$$

and by the same argument

$$\begin{aligned}
&\mathbb{P}\left(\left|\left\{i \in \overline{\text{Top}}_m : \hat{\mu}_i < \mu_i - \varepsilon/2\right\}\right| < m(\varepsilon) - s + k + 1\right) \\
&= \mathbb{P}\left(\left|\left\{i \in \overline{\text{Top}}_m : \hat{\mu}_i \le \mu_i - \varepsilon/2\right\}\right| \ge m(\varepsilon) - s + k\right) \\
&= \mathbb{P}\left(\exists \mathcal{A} \subset \overline{\text{Top}}_m(\varepsilon) : |\mathcal{A}| = m(\varepsilon) - s + k \text{ and } \forall i \in \mathcal{A}, \hat{\mu}_i \ge \mu_i + \varepsilon/2\right) \\
&\le \binom{n - m(\varepsilon)}{s-k} \exp\left(-(s-k)\frac{\varepsilon^2 T}{2n}\right).
\end{aligned}$$

□

C.4 Proof of Theorem 4.17

Proof. Let $\theta = (\varepsilon, \ldots, \varepsilon, 0, \ldots, 0) \in \mathbb{R}^n$ with $\varepsilon > 0$ be a vector of mean observations for each arm where the first m coordinates are nonzero. Let $\hat{\mathcal{S}} \in [n]$ be the output of the algorithm

(recall $|\hat{\mathcal{S}}| = s$). Let $\tilde{M}_\theta$ be the number of false negatives when taking $\hat{\mathcal{S}}$ as the prediction for the true support $[s]$ of θ. Let $\tilde{\mathcal{Q}}_a \subset 2^{[n]}$ be the collection of all subsets of $[n]$ of size s such that $\tilde{M}_\theta = a$. For convenience, let $\gamma = \frac{1}{8}m$. Let $k = \frac{3}{4}s$. We have

$$\begin{aligned}\xi \coloneqq \mathbb{P}_\theta(\tilde{M}_\theta \geq m - k) &= \sum_{a=m-k}^{k} \mathbb{P}_\theta(\tilde{M}_\theta = a) \\ &\geq \sum_{a=3\gamma}^{5\gamma} \mathbb{P}_\theta(\tilde{M}_\theta = a).\end{aligned}$$

Let $\mathsf{first}(\tilde{\mathcal{Q}}_a)$ be the first member of $\tilde{\mathcal{Q}}_a$ in lexicographic order. For example, $\mathsf{first}(\tilde{Q}_a) = [a+1 : a+s]$. Then, by symmetry, one can see that $\mathbb{P}_\theta(\tilde{M}_\theta = a) = |\tilde{\mathcal{Q}}_a|\mathbb{P}_\theta(\hat{\mathcal{S}} = \mathsf{first}(\tilde{\mathcal{Q}}_a))$.

We consider the event on which the 'empirical KL-divergence' concentrates, which by Lemma C.3 holds with probability at least $1 - \delta$:

$$\mathsf{conc}(\theta', \theta) \coloneqq \left\{ \sum_{i=1}^{n}\sum_{t=1}^{B} \ln\left(\frac{p_{\theta'}(X_{i,t}|A_t = i)}{p_\theta(X_{i,t}|A_t = i)}\right) - (1+\rho)B\sum_{i=1}^{n} \mathrm{KL}(\theta'_i, \theta_i) \leq \frac{1}{\rho}\ln(1/\delta) \right\}.$$

We now employ a change of measure argument to $\mathbb{P}_\theta(\hat{\mathcal{S}} = \mathsf{first}(\tilde{\mathcal{Q}}_a))$ and switch θ with θ' that would result in $\tilde{M}_{\theta'} = a - 3\gamma =: b$. For $a \in [3\gamma : 5\gamma]$, this can be achieved with the choice of $\theta' = (0, \ldots, 0, \varepsilon, \ldots, \varepsilon, 0, \ldots, 0)$ that is supported on $[3\gamma + 1 : m + 3\gamma]$. Let Θ be the set of permutations of θ. Then,

$$\begin{aligned}
&\mathbb{P}_\theta(\tilde{M}_\theta = a) \\
&= |\tilde{\mathcal{Q}}_a|\mathbb{P}_\theta(\hat{\mathcal{S}} = \mathsf{first}(\tilde{\mathcal{Q}}_a)) \\
&\geq |\tilde{\mathcal{Q}}_a|\mathbb{P}_{\theta'}(\hat{\mathcal{S}} = \mathsf{first}(\tilde{\mathcal{Q}}_a), \mathsf{conc}(\theta', \theta)) \exp(-(1+\rho)2mB \cdot (\varepsilon^2/2) - \rho^{-1}\ln(1/\delta)) \\
&\qquad\qquad \text{(change of measure; } \rho > 0) \\
&\geq |\tilde{\mathcal{Q}}_a|\mathbb{P}_{\theta'}(\hat{\mathcal{S}} = \mathsf{first}(\tilde{\mathcal{Q}}_a), \cap_{\sigma\in\Sigma^n}\mathsf{conc}(\theta', \sigma(\theta'))) \exp(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln(1/\delta)) \\
&\qquad\qquad \text{(Σ^n: symmetric group of } [n]) \\
&= |\tilde{\mathcal{Q}}_a|\mathbb{P}_\theta(\hat{\mathcal{S}} = \mathsf{first}(\tilde{\mathcal{Q}}_b), \cap_{\sigma\in\Sigma^n}\mathsf{conc}(\theta, \sigma(\theta))) \exp(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln(1/\delta)) \\
&\qquad\qquad \text{(symmetry)} \\
&= \frac{|\tilde{\mathcal{Q}}_a|}{|\tilde{\mathcal{Q}}_b|}\mathbb{P}_\theta(\tilde{M}_\theta = b, \cap_{\sigma\in\Sigma^n}\mathsf{conc}(\theta, \sigma(\theta))) \exp(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln(1/\delta)) \quad \text{(symmetry)} \\
&\geq \frac{|\tilde{\mathcal{Q}}_a|}{|\tilde{\mathcal{Q}}_b|}\left(\mathbb{P}_\theta(\tilde{M}_\theta = b) - |\Theta|\delta\right) \exp\left(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln(1/\delta)\right). \\
&\qquad\qquad \text{($\mathbb{P}(A,B) \geq \mathbb{P}(A) - \mathbb{P}(B^c)$; union bound over } \{\sigma(\theta) : \sigma \in \Sigma^n\})
\end{aligned}$$

Thus,

$$\sum_{a=3\gamma}^{5\gamma} \mathbb{P}_\theta(\tilde{M}_\theta = a)$$
$$\geq \underbrace{\min_{a\in[3\gamma:5\gamma]} \frac{|\tilde{\mathcal{Q}}_a|}{|\tilde{\mathcal{Q}}_{a-3\gamma}|}}_{=:\,Y} \Big(\underbrace{\mathbb{P}_\theta(\tilde{M}_\theta \in [0:2\gamma])}_{\geq 1-\xi} - (2\gamma+1)|\Theta|\delta \Big) \times$$
$$\exp\left(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln(1/\delta)\right).$$

Let us choose $\delta = \frac{1}{2}\cdot\frac{1-\xi}{(2\gamma+1)|\Theta|}$. Since the left hand side above is at most ξ, we have

$$\xi \geq Y\frac{1-\xi}{2}\exp\left(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln\left(\frac{2(2\gamma+1)|\Theta|}{1-\xi}\right)\right).$$

One can consider two cases, namely $\xi \geq \frac{1}{2}$ and $\xi < \frac{1}{2}$, to arrive at

$$\xi \geq \min\left\{\frac{1}{2},\ \exp\left(-(1+\rho)mB\varepsilon^2 - \rho^{-1}\ln\left(4(2\gamma+1)|\Theta|\right) + \ln(4Y)\right)\right\}.$$

It remains to find an appropriate value of ρ. One simple choice is

$$\rho^{-1} = \frac{1}{2}\frac{\ln(4Y)}{\ln\left(4(2\gamma+1)|\Theta|\right)}$$

which satisfies $\rho^{-1} > 0$ as we show later. Thus, we have

$$\xi \geq \min\left\{\frac{1}{2},\ 2\sqrt{Y}\exp\left(-\left(1+2\frac{\ln\left(4(2\gamma+1)|\Theta|\right)}{\ln(4Y)}\right)mB\varepsilon^2\right)\right\}.$$

It remains to figure out bounds for Y and $|\Theta|$. For Y, note that $|\tilde{\mathcal{Q}}_a| = \binom{m}{m-a}\binom{n-m}{s-(m-a)}$. So, for $a \in [3\gamma:5\gamma]$ and $b = a-3\gamma$,

$$\begin{aligned}
\min_a \frac{|\tilde{\mathcal{Q}}_a|}{|\tilde{\mathcal{Q}}_b|} &= \frac{\binom{m}{m-a}\binom{n-m}{s-m+a}}{\binom{m}{m-b}\binom{n-m}{s-m+b}} \\
&= \frac{(m-b)(m-b-1)\cdots(m-a+1)}{a(a-1)\cdots(b+1)} \\
&\quad\times \frac{(n-s-b)(n-s-b-1)\cdots(n-s-a+1)}{(s-m+a)(s-m+a-1)\cdots(s-m+b+1)} \\
&\overset{(a)}{\geq} \left(\frac{m-b}{a}\right)^{a-b}\cdot\left(\frac{n-s-b}{s-m+a}\right)^{a-b} \\
&\geq \left(\frac{6}{5}\right)^{3\gamma}\cdot\left(\frac{n-s-2\gamma}{s-3\gamma}\right)^{3\gamma} \geq \left(\frac{6}{5}\right)^{3\gamma}\cdot\left(\frac{n-s}{s}\right)^{3\gamma} \\
\implies Y = \min_{a\in[3\gamma:5\gamma]} \frac{|\tilde{\mathcal{Q}}_a|}{|\tilde{\mathcal{Q}}_{a-2\gamma}|} &\geq \left(\frac{6}{5}\right)^{3\gamma}\cdot\left(\frac{n-s}{s}\right)^{3\gamma},
\end{aligned}$$

where (a) is due to the fact that $\frac{m-b}{a} > 1$ implies $(m-b-i)/(a-i) \geq (m-b)/a$ for $i \in [0 : a-b-1]$ and, with a similar reasoning, $n \geq s/2 \implies \frac{n-s-a+3\gamma}{s-m+a} > 1 \implies (n-s-b-i)/(s-m+a-i) \geq (n-s-b)/(s-m+a)$. Then, using $|\Theta| = \binom{n}{m} \leq \left(\frac{en}{m}\right)^m$, we have

$$\begin{aligned}\rho = 2\frac{\ln\left(4(2\gamma+1)|\Theta|\right)}{\ln(4Y)} &\leq 2\frac{\ln(m+4) + m\ln\left(\frac{en}{m}\right)}{\ln(4) + \frac{m}{4}\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)} \\ &\overset{(a)}{\leq} 4\frac{m\ln\left(\frac{en}{m}\right)}{\ln(4) + \frac{m}{4}\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)} \\ &\leq 16\frac{\ln(en/m)}{\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)},\end{aligned}$$

where (a) is by $m \leq n/3 \implies \ln(m+4) \leq m\ln(en/m)$.

Altogether,

$$\mathbb{P}_\theta(\tilde{M} \geq \frac{1}{4}m) \geq \min\left\{\frac{1}{2},\ 2\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)^{\frac{3}{16}m}\exp\left(-\left(1+16\cdot\frac{\ln(en/m)}{\ln\left(\frac{6}{5}\cdot\frac{n-s}{s}\right)}\right)mB\varepsilon^2\right)\right\}.$$

To verify that our choice of ρ is nonnegative, it suffices to show that $\frac{6}{5}\cdot\frac{n-s}{s} > 1$, which is true for $s < \frac{6}{11}n$. □

C.5 Proof of Theorem 4.15

As mentioned in the proof sketch we will make use of a bound on the suboptimality of the optimal arm in each phase of elimination, which we prove here first.

Theorem C.4. *For any $\varepsilon \in (0,1)$, the error probability of SH for identifying an ε-good arm satisfies,*

$$\mathbb{P}\left(\mu_{\psi_T} < \mu_1 - \varepsilon\right) \leq 3\log_2 n \cdot \exp\left(-\frac{\varepsilon^2}{32}\frac{T}{n\log_2 n}\right).$$

Proof. To avoid redundancy and for the sake of readability, we assume n is of a power of 2. Let $\varepsilon_1 = \varepsilon/4$, $T' \coloneqq \frac{T}{\log_2 n}$. And define $\varepsilon_{\ell+1} = \frac{3}{4}\cdot\varepsilon_\ell$. For each stage ℓ, define the event G_ℓ as

$$G_\ell \coloneqq \left\{\max_{i\in S_{\ell+1}}\mu_i \geq \max_{i\in S_\ell}\mu_i - \varepsilon_\ell\right\}.$$

Thus as long as $\bigcap_{\ell=1}^{\log_2 n} G_\ell$ happens, we have that the arm returned after the final stage is an ε-good arm, because

$$\sum_{\ell=1}^{\log_2 n} \varepsilon_\ell < \sum_{\ell=1}^{\infty} \left(\frac{3}{4}\right)^{\ell-1} \cdot \varepsilon_1 = \frac{\varepsilon}{4} \sum_{\ell=1}^{\infty} \left(\frac{3}{4}\right)^{\ell-1} \leq \frac{\varepsilon}{4} \lim_{n\to\infty} \frac{1-(3/4)^n}{1-3/4} = \varepsilon.$$

Further, by a union bound,

$$\begin{aligned}\mathbb{P}\left(\mu_{\psi_T} < \mu_1 - \varepsilon\right) \leq & \mathbb{P}\left(\left(\bigcap_{\ell=1}^{\log_2 n} G_\ell\right)^c\right) \\ & \leq \sum_{\ell=1}^{\log_2 n} \mathbb{P}\left(G_\ell^c\right). \end{aligned} \tag{C.14}$$

Let a_ℓ be the best arm in S_ℓ,

$$\begin{aligned}\mathbb{P}\left(G_\ell^c\right) &= \mathbb{P}\left(G_\ell^c, \hat{\mu}_{a_\ell} < \mu_{a_\ell} - \varepsilon_\ell/2\right) + \mathbb{P}\left(G_\ell^c, \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \varepsilon_\ell/2\right) \\ &\leq \mathbb{P}\left(\hat{\mu}_{a_\ell} < \mu_{a_\ell} - \varepsilon_\ell/2\right) + \mathbb{P}\left(G_\ell^c \mid \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \varepsilon_\ell/2\right). \\ &\leq \exp\left(-\frac{\varepsilon_\ell^2}{2}\frac{T'}{|S_\ell|}\right) + \mathbb{P}\left(G_\ell^c \mid \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \varepsilon_\ell/2\right).\end{aligned}$$

For the second term,

$$\begin{aligned}\mathbb{P}\left(G_\ell^c \mid \hat{\mu}_{a_\ell} \geq \mu_{a_\ell} - \varepsilon_\ell/2\right) &\leq \mathbb{P}\left(\left|\{i \in S_\ell \mid \hat{\mu}_i > \mu_i + \varepsilon_\ell/2\}\right| \geq |S_\ell|/2\right) \\ &\overset{(a_1)}{\leq} \frac{\mathbb{E}\left[\left|\{i \in S_\ell \mid \hat{\mu}_i > \mu_i + \varepsilon_\ell/2\}\right|\right]}{|S_\ell|/2} \\ &\leq \frac{|S_\ell| \exp\left(-\frac{\varepsilon_\ell^2}{2}\frac{T'}{|S_\ell|}\right)}{|S_\ell|/2} \\ &= 2\exp\left(-\frac{\varepsilon_\ell^2}{2}\frac{T'}{|S_\ell|}\right).\end{aligned}$$

For (a_1), we use Markov's inequality. Then,

$$\begin{aligned}\mathbb{P}\left(G_\ell^c\right) \leq 3\exp\left(-\frac{\varepsilon_\ell^2}{2}\frac{T'}{|S_\ell|}\right) &= 3\exp\left(-\left(\frac{9}{16}\right)^{\ell-1}\frac{\varepsilon^2}{32}\frac{T'}{2^{-(\ell-1)}n}\right) \\ &= 3\exp\left(-\left(\frac{9}{8}\right)^{\ell-1}\frac{\varepsilon^2}{32}\frac{T'}{n}\right).\end{aligned}$$

Taking the above into (C.14), we have

$$\begin{aligned}\mathbb{P}\left(\mu_{\psi_T} < \mu_1 - \varepsilon\right) &\leq \sum_{\ell=1}^{\log_2 n} \mathbb{P}\left(G_\ell^c\right) \\ &\leq \sum_{\ell=1}^{\log_2 n} 3\exp\left(-\left(\frac{9}{8}\right)^{\ell-1}\frac{\varepsilon^2}{32}\frac{T'}{n}\right) \\ &\leq 3\log_2 n \cdot \exp\left(-\frac{\varepsilon^2}{32}\frac{T}{n\log_2 n}\right).\end{aligned}$$

□

Let us now turn to the proof of Theorem 4.15.

Proof. Let us consider the case where $m \leq n/2$ first.

Let $\ell^* = \log_2 m$, $\varepsilon' = \frac{\varepsilon}{2\log_2 m}$, $T' \coloneqq \frac{T}{\log_2 n}$ and let g_ℓ denote the number of $(m, \ell \cdot \varepsilon')$-good arms at the end of stage $\ell \in [\log_2 m]$, i.e. arms for which $\Delta_i < \Delta_m + \ell \cdot \varepsilon'$. For stage ℓ, we define the event G_ℓ as,

$$G_\ell \coloneqq \left\{ g_\ell \geq 2^{-\ell} \cdot m \right\}.$$

Specifically, we have G_{ℓ^*}, the event that the number of $(m, \varepsilon/2)$-good arms after finishing stage ℓ^* is at least 1. It follows that G_{ℓ^*+1} is the event where the algorithm succeeds in returning an arm in $\text{Top}_m(\varepsilon)$,

$$G_{\ell^*+1} \coloneqq \{\mu_{\psi_T} \geq \mu_m - \varepsilon\}.$$

Then the event $\bigcap_{\ell=1}^{\ell^*+1} G_\ell$ describes a possible sequence of events in each round, ending with the algorithm selecting an arm from $\text{Top}_m(\varepsilon)$. Thus the probability of missing all of the (m, ε)-good arms can be upper bounded as follows. Define the event that one of these good events fails in round ℓ, $F_\ell = \left(\bigcap_{i=1}^{\ell-1} G_i\right) \cap G_\ell^c$ for $\ell > 1$, $F_1 = G_1^c$. Since $\bigcap_{\ell=1}^{\ell^*+1} G_\ell \subset G_{\ell^*+1}$, this implies that $G_{\ell^*+1}^c = \{\mu_{\psi_T} < \mu_m - \varepsilon\} \subset \left(\bigcap_{\ell=1}^{\ell^*+1} G_\ell\right)^c$,

$$\begin{aligned}
\mathbb{P}\left(\mu_{\psi_T} < \mu_m - \varepsilon\right) \leq & \mathbb{P}\left(\left(\bigcap_{\ell=1}^{\ell^*+1} G_\ell\right)^c\right) \\
= & \mathbb{P}\left(\bigcup_{\ell=1}^{\ell^*+1} F_\ell\right) \\
\leq & \sum_{\ell=1}^{\ell^*+1} \mathbb{P}\left(G_\ell^c \mid \bigcap_{i=1}^{\ell-1} G_i\right) && (\mathbb{P}(A,B) \leq \mathbb{P}(A|B)) \\
= & \sum_{\ell=1}^{\ell^*} \mathbb{P}\left(G_\ell^c \mid \bigcap_{i=1}^{\ell-1} G_i\right) + \mathbb{P}\left(G_{\ell^*+1}^c \mid \bigcap_{i=1}^{\ell^*} G_i\right). && \text{(C.15)}
\end{aligned}$$

For the first term of (C.15), we apply the result of Proposition 4.16 to each stage of Sequential Halving, with the parameters therein as $k = 2^{-\ell} \cdot m$, $s = 2^{-\ell} \cdot n$, $m' = 2^{-\ell+1} \cdot m$,

$\varepsilon = \varepsilon'$, and $T = T'$. Thus, for stage ℓ,

$$
\begin{aligned}
&\mathbb{P}\left(G_\ell^c \mid G_{\ell-1}\right) \\
&\le \binom{2^{-\ell+1}\cdot m}{2^{-\ell}\cdot m} \exp\left(-2^{-\ell}\cdot m \frac{\varepsilon'^2 T'}{2\cdot 2^{-(\ell-1)}\cdot n}\right) \\
&\quad + \binom{2^{-\ell+1}\cdot n}{2^{-\ell}\cdot (n-m)} \exp\left(-2^{-\ell}\cdot (n-m) \frac{\varepsilon'^2 T'}{2\cdot 2^{-(\ell-1)}\cdot n}\right) \\
&\overset{(a_3)}{\le} h_{\ell,1} \exp\left(-2^{-\ell}\cdot m \frac{\varepsilon'^2 T'}{2\cdot 2^{-(\ell-1)}\cdot n}\right) + h_{\ell,2} \exp\left(-2^{-\ell}\cdot (n-m) \frac{\varepsilon'^2 T'}{2\cdot 2^{-(\ell-1)}\cdot n}\right) \\
&\le h_{\ell,1} \exp\left(-\frac{m}{2}\frac{\varepsilon'^2 T'}{2n}\right) + h_{\ell,2} \exp\left(-\frac{n-m}{2}\frac{\varepsilon'^2 T'}{2n}\right) \\
&\overset{(a_4)}{\le} h_{\ell,1} \exp\left(-\frac{m}{2}\frac{\varepsilon'^2 T'}{2n}\right) + h_{\ell,2} \exp\left(-\frac{n}{4}\frac{\varepsilon'^2 T'}{2n}\right) \qquad \text{(C.16)} \\
&\le \exp\left(-\frac{m}{2}\frac{\varepsilon^2 T}{2n\log_2^2 m \log_2 n} + \log(2e) 2^{-\ell}\cdot m\right) \\
&\quad + \exp\left(-\frac{n}{4}\frac{\varepsilon^2 T}{2n\log_2^2 m \log_2 n} + \log(4e) 2^{-\ell}\cdot (n-m)\right) \\
&\le \exp\left(-m\left(\frac{\varepsilon^2 T}{4n\log_2^2 m \log_2 n} - 2\log(4e)2^{-\ell}\right)\right) \\
&\quad + \exp\left(-\frac{n}{2}\left(\frac{\varepsilon^2 T}{4n\log_2^2 m \log_2 n} - 2\log(4e)2^{-\ell}\right)\right) \\
&\le 2\exp\left(-m\left(\frac{\varepsilon^2 T}{4n\log_2^2 m \log_2 n} - 2\log(4e)2^{-\ell}\right)\right)
\end{aligned}
$$

where (a_3) is $\binom{2^{-\ell+1}\cdot m}{2^{-\ell}\cdot m} \le (2e)^{2^{-\ell}\cdot m} =: h_{\ell,1}$ (Sterling's formula $\binom{x}{y} \le \left(\frac{ex}{y}\right)^y$) and $\binom{2^{-\ell+1}\cdot n}{2^{-\ell}\cdot (n-m)} \le (2en/(n-m))^{2^{-\ell}\cdot (n-m)} \le (4e)^{2^{-\ell}\cdot n} =: h_{\ell,2}$ and (a_4) is by the assumption $m \le n/2$.

Taking the summation over rounds we have

$$
\sum_{\ell=1}^{\ell^*} \mathbb{P}\left(G_\ell^c \mid G_{\ell-1}\right) \le \text{const}\cdot \log_2 m \cdot \exp\left(-m\left(\frac{\varepsilon^2 T}{4n\log_2^2 m \log_2 n} - \log(4e)\right)\right). \qquad \text{(C.17)}
$$

For the second term of (C.15), $G_{\ell^*+1}^c \mid G_{\ell^*}$ indicates the event where SH fails to return an (m,ε)-good arm given the event that there is at least one $(m,\varepsilon/2)$-good arm after finishing stage ℓ^*. Note that we can consider the stages from ℓ^* to the last stage as a new run of SH with a budget $\left(\log_2 n - \ell^*\right) T' \ge \text{const}\cdot \log_2\left(\frac{n}{m}\right)\frac{T}{\log_2(n)}$. The initial arms are the surviving arms after finishing stage ℓ^*. Note we have $2^{-\ell^*}\cdot n = n/m$ arms surviving, denoted by S_{ℓ^*}, after finishing stage ℓ^*. Let j be the index of the true best arm in S_{ℓ^*}. Conditioning on the event G_{ℓ^*},

we have $\mu_j \geq \mu_m - \frac{\varepsilon}{2}$. Then, by applying the result of Theorem C.4,

$$\begin{aligned}
\mathbb{P}\left(G^{\mathbf{c}}_{\ell^*+1} \mid G_{\ell^*}\right) &= \mathbb{P}\left(\mu_{\psi_T} < \mu_m - \varepsilon \mid G_{\ell^*}\right) \\
&\leq \mathbb{P}\left(\mu_{\psi_T} < \mu_j - \frac{\varepsilon}{2} \mid G_{\ell^*}\right) \\
&\leq 3\log_2\left(\frac{n}{m}\right) \cdot \exp\left(-\frac{m}{128}\frac{\varepsilon^2\left(\log_2 n - \ell^*\right)T'}{n\log_2\left(\frac{n}{m}\right)}\right) \\
&\leq \log_2\left(\frac{n}{m}\right) \cdot \exp\left(-\text{const} \cdot m\frac{\varepsilon^2 T}{n\log_2 n}\right) . \qquad \text{(C.18)}
\end{aligned}$$

Plugging (C.17) and (C.18) into (C.15), we have,

$$\mathbb{P}\left(\mu_{\psi_T} < \mu_m - \varepsilon\right) \leq \log_2 n \cdot \exp\left(-C \cdot m\left(\frac{\varepsilon^2 T}{4n\log_2^2(m)\log_2 n} - \log(4e)\right)\right)$$

for some positive constant $C > 0$. Note the above analysis is for $m > 1$. To incorporate the result of Theorem C.4 for $m = 1$, we rewrite the final formula as

$$\mathbb{P}\left(\mu_{\psi_T} < \mu_m - \varepsilon\right) \leq \log_2 n \cdot \exp\left(-C \cdot m\left(\frac{\varepsilon^2 T}{4n\log_2^2(2m)\log_2 n} - \log(4e)\right)\right)$$

for some positive constant $C > 0$. Thus, there exist a positive constant c_1 such that for $T \geq c_1\frac{(\log(4e)+\log\log n)n\log_2^2(2m)\log_2 n}{\varepsilon^2} = \tilde{\Theta}\left(\frac{n}{\varepsilon^2}\right)$,

$$\mathbb{P}\left(\mu_{\psi_T} < \mu_m - \varepsilon\right) \leq \exp\left(-\tilde{\Theta}\left(m\frac{\varepsilon^2 T}{n}\right)\right).$$

For the case of $m > n/2$, we consider

$$\mathbb{P}\left(\mu_{\psi_T} < \mu_m - \varepsilon\right) \leq \mathbb{P}\left(\mu_{\psi_T} < \mu_{n/2} - \varepsilon\right).$$

Thus, one can repeat the same analysis as above with m replaced by $n/2$. Using $m/2 \leq n/2 \leq m$, the statement of this theorem holds. □