Improved Asymptotics for Multi-armed Bandit Experiments under Optimism-based Policies: Theory and Applications

Anand Kalvit

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy under the Executive Committee of the Graduate School of Arts and Sciences

## COLUMBIA UNIVERSITY

Anand Kalvit
All Rights Reserved


#### Abstract

Improved Asymptotics for Multi-armed Bandit Experiments under Optimism-based Policies: Theory and Applications


## Anand Kalvit

The classical multi-armed bandit paradigm is a foundational framework for online decision making underlying a wide variety of important applications, e.g., clinical trials, advertising, sequential assignments, assortment optimization, etc. This work will examine two salient aspects of decision making that arise naturally in settings with large action spaces. The first issue pertains to the division of samples across arms at the level of a trajectory (or sample-path). Traditional bounds at the ensemble-level (or in expectation) only translate to meaningful pathwise guarantees (high probability bounds) when the separation between mean rewards is "large," commonly referred to as the "well-separated" regime in the literature. On the other hand, applications with a large action space are intrinsically endowed with smaller separations between arm-means (e.g., multiple products of similar quality in e-retail). As a result, classical ensemble-level guarantees for such problems become vacuous at the sample-path level in several settings. This theoretical gap in the understanding of bandit algorithms in the "small gap" regime can be of significant consequence in applications where considerations such as fairness and post hoc inference play an important role. Our work provides the first systematic treatment and analysis of this aspect under the celebrated UCB class of optimism-based bandit algorithms, including a complete diffusion-limit characterization of its regret. The diffusion-scale lens also reveals profound insights and highlights distinctions between UCB and the popular
posterior sampling-based method, Thompson Sampling, such as an "incomplete learning" phenomenon that is characteristic of the latter.

The second research question studied in this work concerns the complexity of decision making in problems where the action space is endowed with a large number of substitutable alternatives. For example, it is common in e-retail for multiple brands to offer similar products (in terms of quality-of-service) that compete for revenue within a given product segment. We model the platform's decision problem in this example as a bandit with countably many arms, and investigate limits of achievable performance under canonical bandit algorithms adapted to this setting. We also propose novel rate-optimal algorithms that leverage results for the "small gap" regime alluded to earlier, and show that these outperform aforementioned conventional adaptations. We extend the countable-armed bandit paradigm to also serve as a basal motif in sequential assignment and dynamic matching problems typical of settings such as online labor markets.

The last chapter of this thesis investigates achievable performance in the countable-armed bandit problem under non-stationarity that is attributable to vanishing arms. This characteristic abstracts away certain attrition and churn processes observable in online markets, e.g., a popular brand may retract its product from a platform owing to under-exposure within its category - a potential negative externality of the exploration carried out by the platform's policy.

## Table of Contents

Acknowledgments ..... viii
Dedication ..... x
Introduction ..... 1
Chapter 1: The classical multi-armed bandit problem: Towards a comprehensive asymp- totic theory ..... 3
1.1 The model and notation ..... 9
1.2 Main results ..... 11
1.2.1 Asymptotics of arm-sampling rates ..... 11
1.2.2 Beyond arm-sampling rates ..... 18
1.3 Numerical experiments ..... 22
1.4 Concluding remarks and open problems ..... 28
Chapter 2: Multi-armed bandits with $K$ arm-types: A new countable-armed formulation ..... 30
2.1 Introduction ..... 30
2.2 Problem formulation ..... 34
2.3 Lower bounds for natural policy classes ..... 35
2.3.1 Achievable performance w.r.t. $\underline{\Delta}$ : Information-theoretic lower bounds ..... 35
2.3.2 Achievable performance w.r.t. $\alpha_{1}$ : A uniform lower bound for front-loading ..... 37
2.3.3 Achievable performance w.r.t. $K$ : The Bandit and the Coupon-collector ..... 38
2.4 Gap and reservoir adaptive policies ..... 39
2.4.1 Explore-then-commit with a pre-specified exploration schedule ..... 40
2.4.2 Explore-then-commit with an adaptive stopping time ..... 42
2.4.3 Towards fully sequential adaptive strategies: Optimism in exploration ..... 46
2.5 Numerical experiments ..... 49
2.6 Concluding remarks, extensions, and open problems ..... 53
Chapter 3: Countable-armed bandits: An application to matching markets ..... 54
3.1 Introduction ..... 54
3.2 Problem formulation ..... 57
3.3 Designing adaptive policies for matching ..... 59
3.3.1 Shifting focus to adaptive sequential sampling strategies tailored to a spe- cific job-type ..... 60
3.3.2 Transferring learning guarantees to the matching problem ..... 60
3.4 Concluding remarks ..... 61
Chapter 4: Countable-armed bandits with dynamic arm-acquisition costs: Towards a non- stationary arm-reservoir ..... 62
4.1 Introduction ..... 62
4.1.1 Related literature ..... 64
4.1.2 Outline of this chapter ..... 66
4.2 Problem formulation ..... 66
4.3 Distribution-agnostic policies ..... 67
4.3.1 A fixed-design policy based on forced exploration ..... 68
4.3.2 A UCB-based approach with adaptive resampling ..... 69
4.4 Natural models of the arm-reservoir ..... 71
4.4.1 Exogenous reservoirs ..... 73
4.4.2 Endogenous reservoirs ..... 75
4.5 Concluding remarks ..... 76
References ..... 78
Appendix A: Appendix to Chapter 1 ..... 84
A. 1 Closed-form expressions for $\lambda_{\rho}^{*}(\theta)$ and $h_{\rho}(\theta)$ ..... 85
A. 2 Auxiliary results ..... 85
A. 3 Proof of Theorem 1 in the "large gap" regime ..... 86
A. 4 Proof of Theorem 1 in the "small gap" regime ..... 86
A.4.1 An almost sure lower bound on the arm-sampling rates ..... 87
A.4.2 Closing the loop ..... 90
A. 5 Proof of Theorem 1 in the "moderate gap" regime ..... 93
A.5.1 Focusing on arm 1 ..... 93
A.5.2 Focusing on arm 2 and concluding ..... 99
A. 6 Proof of Theorem 3 ..... 101
A.6.1 Proof of part (I) ..... 101
A.6.2 Proof of part (II) ..... 103
A. 7 Proof of Theorem 4 ..... 104
A.7.1 The "small gap" regime ..... 104
A.7.2 The "large gap" regime ..... 104
A.7.3 The "moderate gap" regime ..... 105
A. 8 Proof of Theorem 5 ..... 106
A. 9 Proof of Theorem 2 ..... 108
A.9.1 Preamble ..... 108
A.9.2 Proof of part (I) and (II) ..... 113
A. 10 Proof of Fact 2 and Fact 3 ..... 116
A.10.1 Fact 2 ..... 117
A.10.2 Fact 3 ..... 117
Appendix B: Appendix to Chapter 2 ..... 118
B. 1 Proof of Theorem 6 ..... 118
B. 2 Proof of Theorem 7 ..... 121
B. 3 Proof of Theorem 8 ..... 124
B. 4 Proof of Theorem 9 ..... 125
B. 5 Proof of Proposition 1 ..... 131
B. 6 Proof of Theorem 10 ..... 135
B. 7 Auxiliary results used in the analysis of ALG3 ..... 142
B.7. 1 Proof of Lemma 3 ..... 143
B.7.2 Proof of Lemma 4 ..... 144
B. 8 Proof of Theorem 11 ..... 145
B. 9 Auxiliary results used in the analysis of ALG4 ..... 146
B.9.1 Proof of Lemma 5 ..... 147
B.9.2 Proof of Lemma 6 ..... 148
B. 10 Proof of Theorem 12 ..... 148
Appendix C: Appendix to Chapter 4 ..... 151
C. 1 Proof of Theorem 14 ..... 151
C.1. 1 Proof for Model 1 ..... 151
C.1.2 Proof for Model 2 ..... 152
C. 2 Proof of Proposition 2 ..... 153
C. 3 Analysis of Algorithm 4 ..... 156
C.3.1 Proof of Theorem 15 ..... 161
C.3.2 Proof of Theorem 17 ..... 163
C. 4 Auxiliary results used in the analysis of Algorithm 5 ..... 164
C. 5 Analysis of Algorithm 5 ..... 164
C.5.1 Proof of Theorem 16 ..... 167
C.5.2 Proof of Theorem 18 ..... 168

## List of Figures

1.1 Incomplete learning under Thompson Sampling. A two-armed bandit with Bernoulli $(q)$ rewards for each arm: Histograms show the empirical (probability) distribution of $N_{1}(n) / n$ for $n=10,000$ pulls, plotted using 20,000 experiments. Algorithms: UCB1 [10], TS with Beta priors [13] ..... 5
1.2 Failure of CLT under TS. A two-armed bandit with $\Delta=0$ : Arm 1 returns a constant reward of 0.5 , and arm 2 yields rewards distributed as $\operatorname{Bernoulli}(0.5)$. In (a), the histogram shows the empirical (probability) distribution of $N_{1}(n) / n$. Algorithms: TS (Algorithm 2 in [7]) and UCB (UCB1 in [10]). All histograms have $n=10,000$ pulls, and are plotted using $\boldsymbol{\aleph}=20,000$ experiments. ..... 7
$1.3 \boldsymbol{h}_{\boldsymbol{\rho}}(\boldsymbol{\theta})$ vs. $\boldsymbol{\theta}$ for different values of the exploration coefficient $\rho$ in Algorithm 1. The graphs exhibit a unique global maximizer $\theta_{\rho}^{*}$ for each $\rho$. The ordered pairs $\left(\theta_{\rho}^{*}, h_{\rho}\left(\theta_{\rho}^{*}\right)\right)$ for $\rho \in\{1.1,2,3,4\}$ are (1.9, 0.24), (3.5, 0.32), (5.3, 0.39), (7, 0.45). ..... 20
1.4 Diffusion ensemble of UCB: x-axis is normalized time $t \in[0,1]$; y-axis is $\frac{R_{[n t]}^{\pi}}{\sqrt{n}}$ where $n=10,000$ rounds, and $\pi$ is Algorithm 1 with $\rho=2$. ..... 22
1.5 Instance 1: x-axis is time; blue curve plots the evolution of $N_{1}(t) / t$ on a fixed sample-path; red dots indicate the arm pulled at time $t$. ..... 24
1.6 Instance 1: Distribution of regret ..... 24
1.7 Instance 1: Distribution of regret (right $1 \%$ tail) ..... 24
1.8 Instance 1: Allocation statistics for arm 1 ..... 25
1.9 Instance 1: Allocation statistics for arm 1 (left 1\% tail) ..... 25
1.10 Instance 2: x-axis is time; blue curve plots the evolution of $N_{1}(t) / t$ on a fixed sample-path; red dots indicate the arm pulled at time $t$. ..... 25
1.11 Instance 2: Distribution of regret ..... 26
1.12 Instance 2: Distribution of regret (right 1\% tail) ..... 26
1.13 Instance 2: Allocation statistics for arm 1 ..... 26
1.14 Instance 3: Allocation statistics for arm 1 (Histogram shows the empirical dis- tribution of $N_{1}(T) / T$ after $T=10,000$ rounds plotted by simulating $\boldsymbol{N}=10,000$ independent sample-paths) ..... 27
1.15 Instance 3: Posterior evolution under Thompson Sampling (Sample-path\#1) ..... 27
1.16 Instance 3: Posterior evolution under Thompson Sampling (Sample-path\#2) ..... 27
1.17 Instance 3: Diffusion ensemble of UCB and Uniform Sampling (fair coin toss at each step); x -axis is normalized time $t \in[0,1] ; \mathrm{y}$-axis is $\frac{R_{[t T]}^{\pi}}{\sqrt{T}}$ where $T=10,000$ rounds ..... 28
1.18 Instance 2 (Cross-sectional statistics): Histograms show the empirical distributionof $\left(R_{T}^{\pi} / \sqrt{T}-c / 2\right) / \sqrt{\frac{\sigma_{1}^{2}+\sigma_{2}^{2}}{2}}$ where $T=10,000$ rounds, $c=1$, and $\sigma_{1}^{2}, \sigma_{2}^{2}$ are thereward variances.28
2.1 $K=2$ and $\boldsymbol{\alpha}=(1 / 2,1 / 2)$ : Achievable regret in 2-CAB vis-à-vis 2-MAB. ..... 52
$2.2 K=2$ and $\alpha=(1 / 2,1 / 2)$ : An instance with Bernoulli $0.6,0.4$ rewards. ..... 52
$2.3 \quad K=2$ and $\boldsymbol{\alpha}=(1 / 2,1 / 2)$ : An instance with Bernoulli $0.9,0.5$ rewards. ..... 52
$2.4 \quad K=3$ and $\boldsymbol{\alpha}=(1 / 3,1 / 3,1 / 3)$ : Instance with Bernoulli $0.9,0.5,0.1$ rewards. ..... 52
3.1 Possible distribution of worker-types for $\mathcal{J}=\{\mathbf{1}, \mathbf{2}\}$ and $\boldsymbol{K}_{\mathbf{1}}=\boldsymbol{K}_{\mathbf{2}}=\mathbf{2}$. The darker shades represent type 1 (optimal) workers while the lighter shades represent type 2 (inferior) workers w.r.t. each job-type in $\mathcal{J}$. In this example, no worker can simultaneously be optimal for both job-types.57

## Acknowledgements

I had the singular honor and privilege of being amidst some exceptional people in the past several years. I owe a big part of what and where I am today to each one of you. To Assaf first and foremost, thank you for your unreserved support throughout the years. The creative liberty you gave me has been the cornerstone of my research, and this thesis would simply not exist without it. You have been a great advisor, mentor, role model and friend, and I am fortunate for the opportunity to learn so much from you during this time.

To Yash, thank you for introducing me to the ideas, topics and papers that would eventually shape my interests and guide my segue into the world of bandits. Thanks a lot also for your prompt support from time to time.

To Omar, Dan, and Hongyao, many thanks for your advice and support on a great many things and for always being available. Your perspectives have helped me immensely get clarity on important decisions.

To Carri and Jing, thank you for being incredibly supportive as coordinator of the PhD program. To everyone else at DRO, faculty and staff, thank you for being ever so generous with time, and gracious with help and advice over the years.

To my friends and colleagues in New York City, you have been a wonderful company and source of inspiration. You made my time here memorable. I have been lucky to forge some extraordinary friendships and bonds during this period for which I will always remain thankful.

To my gurus at IIT Bombay, I will be eternally grateful for your support and role in guiding me to this day. To all my teachers in life, thank you for believing in me.

Lastly, none of this could have been possible without the unconditional love, support and trust of my closest friends and family (including and especially the furry ones). You have had my back through thick and thin. Thank you for being my rock.

Anand
March 2023, New York

## Dedication

To my teachers and family

## Introduction

This thesis conducts inquiry into the theoretical underpinnings of the celebrated multi-armed bandit paradigm with a two-fold objective: (i) advancing the frontier of knowledge for classical algorithms; and (ii) distilling insights to guide algorithm design for broader problems involving sequential decision making under parameter uncertainty. Forthcoming paragraphs briefly elucidate the contributions in individual chapters of this thesis.

Chapter 1 provides new results on the arm-sampling behavior of popular algorithms for the stochastic multi-armed bandit problem, such as UCB and Thompson Sampling, leading to several important insights. Among these, it is shown that arm-sampling rates under UCB are asymptotically deterministic regardless of the problem complexity; this discovery facilitates new sharp asymptotics as well as a novel alternative proof of the algorithm's worst-case regret. The chapter also provides the first complete process-level characterization of the multi-armed bandit problem under UCB in the conventional diffusion limit. The diffusion limit, among other things, reveals profound distinctions between UCB and Thompson Sampling that have significant implications for areas such as adaptive inference and algorithmic fairness.

Chapter 2 introduces a countably many-armed bandit problem motivated by sequential stochastic assignments in large markets, and proposes a class of online adaptive policies that achieve rate-optimal regret. The design and analysis of these policies is facilitated in part by the results and technical machinery developed in Chapter 1. It is also established that absent these refinements, conventional bandit policies adapted to this problem setting are inferior in a precise sense. The countable-armed bandit model also encapsulates key elements of several applications
with a greater degree of complexity and provides a tractable basal motif for their analysis; notable examples are the design of personalized recommender systems, and matching algorithms for online labor markets.

Chapter 3 discusses a stylized application of the countable-armed problem to online labor markets where a centralized planner must match "jobs" to "workers" dynamically subject to uncertainty about arrivals, preferences, skills and population-level distributions thereof.

Chapter 4 lifts the aforementioned countable-armed bandit model to an incentive-driven non-stationary setting where arms may potentially "vanish" over time. The vanishing arms characteristic is modeled after phenomena such as customer disengagement that are widely reported in online markets serving a large population of strategic agents.

Proofs and auxiliary technical results are relegated to the appendix.

# Chapter 1: The classical multi-armed bandit problem: Towards a comprehensive asymptotic theory 

Background and motivation. The multi-armed bandit (MAB) paradigm provides a succinct abstraction of the quintessential exploration vs. exploitation trade-offs inherent in many sequential decision making problems. This has origins in clinical trial studies dating back to [1] which gave rise to the earliest known MAB heuristic, Thompson Sampling. Today, the MAB problem manifests itself in various forms with applications ranging from dynamic pricing and online auctions to packet routing, scheduling, e-commerce and matching markets among others (see [2] for a comprehensive survey of different formulations). In the canonical stochastic MAB problem, a decision maker (DM) pulls one of $K$ arms sequentially at each time $t \in\{1,2, \ldots\}$, and receives a random payoff drawn according to an arm-dependent distribution. The DM, oblivious to the statistical properties of the arms, must balance exploring new arms and exploiting the best arm played thus far in order to maximize her cumulative payoff over the horizon of play. This objective is equivalent to minimizing the regret relative to an oracle with perfect ex ante knowledge of the optimal arm (the one with the highest mean reward). The classical stochastic MAB problem is fully specified by the tuple $\left(\left(\mathcal{P}_{i}\right)_{1 \leqslant i \leqslant K}, n\right)$, where $\mathcal{P}_{i}$ denotes the distribution of rewards associated with the $i^{\text {th }}$ arm, and $n$ the horizon of play.

The statistical complexity of regret minimization in the stochastic MAB problem is governed by a key primitive called the gap, denoted by $\Delta$, which accounts for the difference between the top two arm mean rewards in the problem. For a "well-separated" or "large gap" instance, i.e., a fixed $\Delta$ bounded away from 0 , the seminal paper of [3] showed that the order of the smallest achievable regret is logarithmic in the horizon. There has been a plethora of subsequent work involving algorithms which can be fine-tuned to achieve a regret arbitrarily close to the optimal
rate discovered in aforementioned paper (see [4, 5, 6, 7, 8], etc., for a few notable examples). On the other hand, no algorithm can achieve an expected regret smaller than $C \sqrt{n}$ for a fixed $n$ (the constant hides dependence on the number of arms) uniformly over all problem instances (also called minimax regret); see, e.g., [9], Chapter 15. The saddle-point in this minimax formulation occurs at a gap that satisfies $\Delta \asymp 1 / \sqrt{n}$. This has a natural interpretation: approximately $1 / \Delta^{2}$ samples are required to distinguish between two distributions with means separated by $\Delta$; at the $1 / \sqrt{n}$-scale, it becomes statistically impossible to distinguish between samples from the top two arms within $n$ rounds of play. If the gap is smaller, despite the increased difficulty in the hypothesis test, the problem becomes "easier" from a regret perspective. Thus, $\Delta \asymp 1 / \sqrt{n}$ is the statistically "hardest" scale for regret minimization. A number of popular algorithms achieve the $\sqrt{n}$ minimaxoptimal rate (modulo constants), see, e.g., [8, 7], and many more do this within poly-logarithmic factors in $n$. Many of these are variations of the celebrated upper confidence bound algorithms, e.g., UCB1 [10], that achieve a minimax regret of $O(\sqrt{n \log n})$, and at the same time also deliver an instance-optimal regret of $O(\log n)$ (modulo constant multiplicative factors).

A major driver of the regret performance of an algorithm is its arm-sampling characteristics. For example, in the instance-dependent (large gap) setting, optimal regret guarantees imply that the fraction of time the optimal arm(s) are played approaches 1 in probability, as $n$ grows large. However, this fails to provide any meaningful insights as to the distribution of arm-pulls for smaller gaps, e.g., the $\Delta \asymp 1 / \sqrt{n}$ "small gap" that governs the "worst-case" instance-independent setting.

An illustrative numerical example involving "small gap." Consider an A/B testing problem (e.g., a vaccine clinical trial) where the experimenter is faced with two competing objectives: first, to estimate the efficacy of each alternative with the best possible precision given a budget of samples, and second, keeping the overall cost of the experiment low. This is a fundamentally hard task and algorithms incurring a low cumulative cost typically spend little time exploring suboptimal alternatives, resulting in a degraded estimation precision (see, e.g., [11]). In other words, algorithms tailored for (cumulative) regret minimization may lack statistical power [12]. While this trade-off is unavoidable in "well-separated" instances, numerical evidence suggests a plausible
resolution in instances with "small" gaps as illustrated below. For example, such a situation might arise in trials conducted using two similarly efficacious vaccines (abstracted away as $\Delta \approx 0$ ). To illustrate the point more vividly, consider the case where $\Delta$ is exactly 0 (of course, this information is not known to the experimenter). This setting is numerically illustrated in Figure 1.1, which shows the empirical distribution of $N_{1}(n) / n$ (the fraction of time arm 1 is played until time $n$ ) in a two-armed bandit with $\Delta=0$, under two different algorithms (UCB and Thompson Sampling), and two different reward configurations.


Figure 1.1: Incomplete learning under Thompson Sampling. A two-armed bandit with Bernoulli $(q)$ rewards for each arm: Histograms show the empirical (probability) distribution of $N_{1}(n) / n$ for $n=10,000$ pulls, plotted using 20,000 experiments. Algorithms: UCB1 [10], TS with Beta priors [13].

A desirable property of the outcome in this setting is to have a linear allocation of the sampling budget per arm on almost every sample-path of the algorithm, as this leads to "complete learning:" an algorithm's ability to discern statistical indistinguishability of the arm-means, and induce a "balanced" allocation in that event. However, despite the simplicity of the zero-gap scenario, it is far from obvious whether the aforementioned property may be satisfied for standard bandit algorithms such as UCB and Thompson Sampling. Indeed, Figure 1.1 exhibits a striking difference between the two. The concentration around $1 / 2$ observable in Figure 1.1(a) indicates that UCB results in an approximately "balanced" sample-split, i.e., the allocation is roughly $n / 2$ per arm for large $n$ (and this is observed for "most" sample-paths). In fact, we will later see that the "bell curve" in Figure 1.1(a) eventually collapses into the Dirac measure at $1 / 2$ (Theorem 1).

On the other hand, under Thompson Sampling, the allocation of samples across arms may be arbitrarily "imbalanced" despite the arms being statistically identical, as seen in Figure 1.1(b) (see, for contrast, Figure 1.1(c), where the allocation is perfectly "balanced"). Namely, the distribution of the posterior may be such that arm 1 is allocated anywhere from almost no sampling effort all the way to receiving almost the entire sampling budget, as Figure 1.1(b) suggests. Non-degeneracy of arm-sampling rates is observable also under the more widely used version of the algorithm that is based on Gaussian priors and Gaussian likelihoods (Algorithm 2 in [7]); see Figure 1.2(a). Such behavior can be detrimental for ex post causal inference in the general $\mathrm{A} / \mathrm{B}$ testing context, and the vaccine testing problem referenced earlier. This is demonstrated via an instructional example of a two-armed bandit with one deterministic reference arm (aka the "one-armed" bandit paradigm), illustrated in Figure 1.2, and discussed below.

A numerical example illustrating inference implications. Consider a model where arm 1 returns a constant reward of 0.5 , while arm 2 yields rewards distributed as Bernoulli(0.5). In this setup, the estimate of the gap $\Delta$ (average treatment effect in causal inference parlance) after $n$ rounds of play is given by $\hat{\Delta}=\bar{X}_{2}(n)-0.5$, where $\bar{X}_{2}(n)$ denotes the empirical mean reward of arm 2 at time $n$. The $\mathcal{Z}$ statistic associated with this gap estimator is given by $\mathcal{Z}=2 \sqrt{N_{2}(n)} \hat{\Delta}$, where $N_{2}(n)$ is the visitation count of arm 2 at time $n$. In the absence of any sample-adaptivity in the arm 2 data, results from classical statistics such as the Central Limit Theorem (CLT) would posit an asymptotically Normal distribution for $\mathcal{Z}$. However, since the algorithms that play the arms are adaptive in nature, e.g., UCB and Thompson Sampling, asymptotic-normality may no longer be guaranteed. Indeed, the numerical evidence in Figure 1.2(b) strongly points to a significant departure from asymptotic-normality of the $\mathcal{Z}$ statistic associated with the gap estimator under Thompson Sampling (TS). Non-normality of the $\mathcal{Z}$ statistic can be problematic for inferential tasks, e.g., it can lead to statistically unsupported inferences in the binary hypothesis test $\mathcal{H}_{0}: \Delta=0$ vs. $\mathcal{H}_{1}: \Delta \neq 0$ performed using confidence intervals constructed as per the conventional CLT approximation. In sharp contrast, our work shows that UCB satisfies a certain "balanced" sampling property (such as that in Figure 1.1(a)) in instances with "small" gaps, formally
stated as Theorem 1, that drives the $\mathcal{Z}$ statistic towards asymptotic-normality in the aforementioned binary hypothesis testing example (asymptotic-normality being a consequence of Theorem 5). Furthermore, since the $\sqrt{n}$-normalized "stochastic" regret (defined in (1.1) in §1.1) equals $-\left(\sqrt{N_{2}(n) /(4 n)}\right) \mathcal{Z}$, it follows that this too, satisfies asymptotic-normality under UCB (due to Theorem 5, in conjunction with Theorem 1). These properties are evident in Figure 1.2(c) below, and signal reliability of ex post causal inference (under classical assumptions like validity of CLT) from "small gap" data collected by UCB vis-à-vis Thompson Sampling (TS). The reliability of inference under TS may be doubtful even in the limit of infinite data, as Figure 1.2(b) suggests.


Figure 1.2: Failure of CLT under TS. A two-armed bandit with $\Delta=0$ : Arm 1 returns a constant reward of 0.5 , and arm 2 yields rewards distributed as $\operatorname{Bernoulli}(0.5)$. In (a), the histogram shows the empirical (probability) distribution of $N_{1}(n) / n$. Algorithms: TS (Algorithm 2 in [7]) and UCB (UCB1 in [10]). All histograms have $n=10,000$ pulls, and are plotted using $\boldsymbol{N}=20,000$ experiments.

While traditional literature on stochastic bandits is dedicated primarily to the regret minimization problem, there has been significant recent interest also in finer-grain properties of popular "adaptive" MAB algorithms such as UCB and Thompson Sampling. For example, a recent line of work ( $[14,15,16])$ investigates the "bias" of optimistic algorithms like UCB. The focus of our work is on understanding the distribution of arm-pulls, which as discussed earlier, has significant bearings on ex post causal inference from data collected adaptively by bandit algorithms (see, e.g., $[17,18,19]$, etc., and references therein for recent developments), algorithmic fairness in the broader context of fairness in machine learning (see [20] for a survey), as well as on novel formulations of the MAB problem such as [21]. Below, we discuss extant literature relevant to our
line of work.
Previous work. The study of "well-separated" instances, or the large gap regime, is supported by rich literature. For example, [4] provides high-probability bounds on arm-sampling rates under a parametric family of UCB algorithms. However, as the gap diminishes, leading to the so called small gap regime, the aforementioned bounds become vacuous. The understanding of arm-sampling behavior remains relatively under-studied here even for popular algorithms such as UCB and Thompson Sampling. This regime is of special interest in that it also covers the classical diffusion scaling ${ }^{1}$, where $\Delta \asymp 1 / \sqrt{n}$, which as discussed earlier, corresponds to instances that statistically constitute the "worst-case" for hypothesis testing and regret minimization. Recently, a partial diffusion-limit characterization of the arm-sampling distribution under a version of Thompson Sampling with horizon-dependent prior variances ${ }^{2}$ was provided in [19] as a solution to a certain stochastic differential equation (SDE). The numerical solution to said SDE was observed to have a non-degenerate distribution on $[0,1]$. Similar numerical observations on non-degeneracy of the arm-sampling distribution also under standard versions of Thompson Sampling were reported in $[23,21]$, among others, albeit limited only to the special case of $\Delta=0$, and absent a theoretical explanation for the aforementioned observations. Thus, outside of the so called "easy" problems, where $\Delta$ is bounded away from 0 by an absolute constant, theoretical understanding of the sampling behavior of bandit algorithms remains an open area of research.

Contributions. In this paper, we provide the first complete asymptotic characterization of armsampling distributions under canonical UCB (Algorithm 1) as a function of the gap $\Delta$ (Theorem 1). This gives rise to a fundamental insight: arm-sampling rates are asymptotically deterministic under UCB regardless of the hardness of the instance. We also provide the first theoretical explanation for an "incomplete learning" phenomenon under Thompson Sampling (Algorithm 3) alluded to in Figure 1.1, as well as a sharp dichotomy between Thompson Sampling and UCB evident therein (Theorem 3). This result earmarks an "instability" of Thompson Sampling in terms of the limiting

[^0]arm-sampling distribution. As a sequel to Theorem 1, we provide the first complete characterization of the worst-case performance of canonical UCB (Theorem 4). One consequence is that the $O(\sqrt{n \log n})$ minimax regret of UCB is strictly unimprovable in a precise sense. Moreover, our work also leads to the first process-level characterization of the two-armed bandit problem under canonical UCB in the classical diffusion limit, according to which a suitably normalized cumulative reward process converges in law to a Brownian motion with fully characterized drift and infinitesimal variance (Theorem 5). To the best of our knowledge, this is the first such characterization of UCB-type algorithms. Theorem 5 facilitates a complete distribution-level characterization of UCB's diffusion-limit regret, thereby providing sharp insights as to the problem's minimax complexity. Such distribution-level information may also be useful for a variety of inferential tasks, e.g., construction of confidence intervals (see the binary hypothesis testing example referenced in Figure 1.2(c)), among others. We believe our results may also present new design considerations, in particular, how to achieve, loosely speaking, the "best of both worlds" for Thompson Sampling, by addressing its "small gap" instability. Lastly, we note that our proof techniques are markedly different from the conventional methodology adopted in MAB literature ( $[4,2,7]$ ), and may be of independent interest in the study of related learning algorithms.

Organization of the chapter. A formal description of the model and the canonical UCB algorithm is provided in §1.1. All theoretical propositions are stated in §1.2, along with a highlevel overview of their scope and proof sketch; detailed proofs and ancillary results are relegated to Appendix A. Finally, concluding remarks and open problems are presented in §1.4.

### 1.1 The model and notation

The technical development in this paper will focus on the two-armed problem purely for expositional reasons. The restriction to two-armed bandits has precedence also in the literature due to its tractability for sharp asymptotic analyses, see, e.g., [24]. This setting encapsulates the core statistical complexity of the MAB problem in the "small gap" regime, as well as concisely highlighting the key novelties in our approach. Before describing the model formally, we introduce the
following asymptotic conventions.
Notation. We say $f(n)=o(g(n))$ or $g(n)=\omega(f(n))$ if $\lim _{n \rightarrow \infty} \frac{f(n)}{g(n)}=0$. Similarly, $f(n)=$ $\mathcal{O}(g(n))$ or $g(n)=\Omega(f(n))$ if $\lim \sup _{n \rightarrow \infty}\left|\frac{f(n)}{g(n)}\right| \leqslant C$ for some constant $C$. If $f(n)=O((g(n)))$ and $f(n)=\Omega((g(n)))$ hold simultaneously, we say $f(n)=\Theta(g(n))$, or $f(n) \asymp g(n)$, and we write $f(n) \sim g(n)$ in the special case where $\lim _{n \rightarrow \infty} \frac{f(n)}{g(n)}=1$. If either sequence $f(n)$ or $g(n)$ is random, and one of the aforementioned ratio conditions holds in probability, we use the subscript $p$ with the corresponding Landau symbol. For example, $f(n)=o_{p}(g(n))$ if $f(n) / g(n) \xrightarrow{p} 0$ as $n \rightarrow \infty$. Lastly, the notation ' $\Rightarrow$ ' will be used for weak convergence.

The model. The arms are indexed by $\{1,2\}$. Each arm $i \in\{1,2\}$ is characterized by a reward distribution $\mathcal{P}_{i}$ supported on $[0,1]$ with mean $\mu_{i}$. The difference between the two mean rewards, aka the gap, is given by $\Delta=\left|\mu_{1}-\mu_{2}\right|$; as discussed earlier, this captures the hardness of an instance. The sequence of rewards associated with the first $m$ pulls of arm $i$ is denoted by $\left(X_{i, j}\right)_{1 \leqslant j \leqslant m}$. The rewards are assumed to be i.i.d. in time, and independent across arms. ${ }^{3}$ The number of pulls of arm $i$ up to (and including) time $t$ is denoted by $N_{i}(t)$. A policy $\pi:=\left(\pi_{t}\right)_{t \in \mathbb{N}}$ is an adapted sequence that prescribes pulling an $\operatorname{arm} \pi_{t} \in \mathcal{S}$ at time $t$, where $\mathcal{S}$ denotes the probability simplex on $\{1,2\}$. The natural filtration at time $t$ is given by $\mathcal{F}_{t}:=\sigma\left\{\left(\pi_{s}\right)_{s \leqslant t},\left(\left(X_{i, j}\right)_{j \leqslant N_{i}(t)}: i=1,2\right)\right\}$. The stochastic regret of policy $\pi$ after $n$ plays, denoted by $R_{n}^{\pi}$, is given by

$$
\begin{equation*}
R_{n}^{\pi}:=\sum_{t=1}^{n}\left[\max \left(\mu_{1}, \mu_{2}\right)-X_{\pi_{t}, N_{\pi_{t}}(t)}\right] . \tag{1.1}
\end{equation*}
$$

The decision maker is interested in the problem of minimizing the expected regret, given by $\inf _{\pi \in \Pi} \mathbb{E} R_{n}^{\pi}$, where $\Pi$ is the set of policies satisfying the non-anticipation property $\pi_{t}: \mathcal{F}_{t-1} \rightarrow$ $\mathcal{S}, 1 \leqslant t \leqslant n$, and the expectation is w.r.t. the randomness in reward realizations as well as possible randomness in the policy $\pi$. In this paper, we will focus primarily on the canonical UCB policy given by Algorithm 1 below. This policy is parameterized by an exploration coefficient $\rho$, which controls its arm-exploring rate. The standard UCB1 policy [10] corresponds to Algorithm 1 with $\rho=2$; the effect of $\rho$ on the expected and high-probability regret bounds of the algorithm is

[^1]well-documented in [4] for problems with a "large gap." In what follows, $\bar{X}_{i}(t-1)$ denotes the empirical mean reward from arm $i \in\{1,2\}$ at time $t-1$, i.e., $\bar{X}_{i}(t-1):=\sum_{j=1}^{N_{i}(t-1)} X_{i, j} / N_{i}(t-1)$.

```
Algorithm 1 The canonical UCB policy for two-armed bandits.
    Input: Exploration coefficient \(\rho \in \mathbb{R}_{+}\).
    At \(t=1,2\), play each arm \(i \in\{1,2\}\) once.
    for \(t \in\{3,4, \ldots\}\) do
4: \(\quad\) Play arm \(\pi_{t} \in \arg \max _{i \in\{1,2\}}\left(\bar{X}_{i}(t-1)+\sqrt{\frac{\rho \log (t-1)}{N_{i}(t-1)}}\right)\).
```


### 1.2 Main results

Algorithm 1 is known to achieve $\mathbb{E} R_{n}^{\pi}=O(\log n)$ in the instance-dependent setting, and $\mathbb{E} R_{n}^{\pi}=O(\sqrt{n \log n})$ in the "small gap" minimax setting. The primary focus of this paper is on the distribution of arm-sampling rates, i.e., $N_{i}(n) / n, i \in\{1,2\}$. Our main results are split across two sub-sections; §1.2.1 examines the behavior of UCB (Algorithm 1) as well as another popular bandit algorithm, Thompson Sampling (specified in Algorithm 3). §1.2.2 is dedicated to results on the (stochastic) regret of Algorithm 1 under the $\Delta \asymp \sqrt{(\log n) / n}$ "worst-case" gap and the $\Delta \asymp 1 / \sqrt{n}$ "diffusion-scaled" gap. We will slightly overload notation by adding the subscript $n$ to $\Delta$ (leading to $\Delta_{n}$ ) in order to clearly highlight its dependence on the horizon $n$. We reemphasize that this paper is focused on the setting where $\Delta_{n}$ scales with $n$; reward variances remain invariant w.r.t. $n$.

### 1.2.1 Asymptotics of arm-sampling rates

Theorem 1 (Arm-sampling rates under UCB in 2-MAB) Let $i^{*} \in \arg \max \left\{\mu_{i}: i=1,2\right\}$ with ties broken arbitrarily. Then, the following results hold for arm $i^{*}$ as $n \rightarrow \infty$ under Algorithm 1 with $\rho>1$ :
(I) "Large gap:" If $\Delta_{n}=\omega\left(\sqrt{\frac{\log n}{n}}\right)$, then

$$
\frac{N_{i^{*}}(n)}{n} \xrightarrow{p} 1 .
$$

(II) "Small gap:" If $\Delta_{n}=o\left(\sqrt{\frac{\log n}{n}}\right)$, then

$$
\frac{N_{i^{*}}(n)}{n} \xrightarrow{p} \frac{1}{2} .
$$

(III) "Moderate gap:" If $\Delta_{n} \sim \sqrt{\frac{\theta \log n}{n}}$ for some fixed $\theta \geqslant 0$, then $N_{i^{*}}(n) / n \xrightarrow{p} \lambda_{\rho}^{*}(\theta)$, where the limit is the unique solution (in $\lambda$ ) to

$$
\begin{equation*}
\frac{1}{\sqrt{1-\lambda}}-\frac{1}{\sqrt{\lambda}}=\sqrt{\frac{\theta}{\rho}} \tag{1.2}
\end{equation*}
$$

and is monotone increasing in $\theta$, with $\lambda_{\rho}^{*}(0)=1 / 2$ and $\lambda_{\rho}^{*}(\theta) \rightarrow 1$ as $\theta \rightarrow \infty$.
Remark 1 (Permissible values of $\rho$ in Algorithm 1) For $\rho>1$, the expected regret of the policy $\pi$ given by Algorithm 1 is bounded as $\mathbb{E} R_{n}^{\pi} \leqslant C \rho\left(\frac{\log n}{\Delta}+\frac{\Delta}{\rho-1}\right)$ for some absolute constant $C>0$; the upper bound becomes vacuous for $\rho \leqslant 1$ (see [4], Theorem 7). We therefore restrict Theorem 1 to $\rho>1$ to ensure that $\mathbb{E} R_{n}^{\pi}$ remains non-trivially bounded for all $\Delta$.

Discussion and intuition. Theorem 1 essentially asserts that the sampling rates $N_{i}(n) / n$, $i \in\{1,2\}$ are asymptotically deterministic in probability under canonical UCB; $\Delta$ only serves to determine the value of the limiting constant. The "moderate" gap regime offers a continuous interpolation from instances with zero gaps to instances with "large" gaps as $\theta$ sweeps over $\mathbb{R}_{+}$in that $\lambda_{\rho}^{*}(\theta)$ increases monotonically from $1 / 2$ at $\theta=0$ to 1 at $\theta=\infty$, consistent with intuition. The special case of $\theta=0$ is numerically illustrated in Figure 1.1(a). The tails of $N_{i^{*}}(n) / n$ decay polynomially fast near the end points of the interval $[0,1]$ with the best possible rate approaching $O\left(n^{-3}\right)$, occurring for $\theta=0$. However, as $N_{i^{*}}(n) / n$ approaches its limit, convergence becomes slower and is dominated by fatter $\Theta\left(\sqrt{\frac{\log \log n}{\log n}}\right)$ tails. The behavior in this regime is regulated by the $O(\sqrt{n \log \log n})$ envelope of the zero-drift random walk process driving the algorithm's regret (see proof of Theorem 1 for details).

Proof sketch. To provide the most intuitive explanation, we pivot to the special case where the arms have identical reward distributions, and in particular, $\Delta=0$. The natural candidate then
for the limit of the empirical sampling rate is $1 / 2$. On a high level, the proof relies on polynomially decaying bounds in $n$ for $\epsilon$-deviations of the form $\mathbb{P}\left(\left|\frac{N_{1}(n)}{n}-\frac{1}{2}\right| \geqslant \epsilon\right)$ derived using the standard trick for bounding the number of pulls of any arm on a given sample-path, to wit, for any $u, n \in \mathbb{N}, N_{1}(n)$ can be bounded above by $u+\sum_{t=u+1}^{n} \mathbb{1}\left\{\pi_{t}=1, N_{1}(t-1) \geqslant u\right\}$, path-wise. Setting $u=\lceil(1 / 2+\epsilon) n\rceil$ in this expression, one can subsequently show via an analysis involving careful use of the policy structure together with appropriate Chernoff bounds that with high probability (approaching 1 as $n \rightarrow \infty$ ), $N_{1}(n) / n \leqslant 1 / 2+\varepsilon_{\rho}$ for some $\varepsilon_{\rho} \in(0,1 / 2)$ that depends only on $\rho$. An identical result would naturally hold also for the other arm by symmetry arguments, and therefore we arrive at a meta-conclusion that $N_{i}(n) / n \geqslant 1 / 2-\varepsilon_{\rho}>0$ for both arms $i \in\{1,2\}$ with high probability (approaching 1 as $n \rightarrow \infty$ ). It is noteworthy that said conclusion cannot be arrived at for an arbitrary $\epsilon>0$ (in place of $\varepsilon_{\rho}$ ) since the polynomial upper bounds on $\mathbb{P}\left(\left|\frac{N_{1}(n)}{n}-\frac{1}{2}\right| \geqslant \epsilon\right)$ derived using the aforementioned path-wise upper bound on $N_{1}(n)$, become vacuous if $u$ is set "too close" to $n / 2$, i.e., if $\epsilon$ is "near" 0 . Extension to the full generality of $\epsilon>0$ is achieved via a refined analysis that uses the Law of the Iterated Logarithm (see [25], Theorem 8.5.2), together with the previous meta-conclusion, to obtain fatter $O\left(\sqrt{\frac{\log \log n}{\log n}}\right)$ tail bounds when $\epsilon$ is near 0 . Here, it is imperative to point out that the " $\log n$ " appearing in the denominator is essentially from the $\sqrt{\rho \log t}$ optimistic bias term of UCB (see Algorithm 1), and therefore the convergence will, as such, hold also for other variations of the policy that have "less aggressive" $\omega(\log \log t)$ exploration functions vis-à-vis $\log t$. However, this will be achieved at the expense of the policy's expected regret performance, as noted in Remark 1. We also note that the extremely slow $O\left(\sqrt{\frac{\log \log n}{\log n}}\right)$ convergence is not an artifact of our analysis, but in fact, supported by the numerical evidence in Figure 1.1(a), suggestive of a plausible non-convergence (to $1 / 2$ ) in the limit. We believe such observations in previous works likely led to incorrect folk conjectures ruling out the existence of a deterministic limit under UCB à la Theorem 1 (see, e.g., [23] and references therein). The proof for a general $\Delta$ in the "small" and "moderate" gap regimes is skeletally similar to that for $\Delta=0$, albeit guessing a candidate limit for $N_{i^{*}}(n) / n$ is non-trivial; a closed-form expression for $\lambda_{\rho}^{*}(\theta)$ is provided in Appendix A.1. Full details of the proof of Theorem 1 are provided in §A.3,A.4,A. 5

## Remark 2 (Possible generalizations of Theorem 1) 1. The behavior of UCB policies is largely

 governed by their optimistic bias. While Theorem 1 only covers the generic UCB policy with $\sqrt{\rho \log t}$ bias, results of the form $N_{i}(n) / n \xrightarrow{p} c_{i}$ for some $c_{i} \in(0,1)$ continue to hold also under smaller $\omega(\sqrt{\rho \log \log t})$ bias (driven by the Law of the Iterated Logarithm). We believe this observation will be useful when examining more complicated UCB-inspired policies such as KL-UCB [6], DMED [26], etc.2. A simple extension to the $K$-armed setting is provided in Theorem 2 below.
```
Algorithm 2 The canonical UCB policy for \(K\)-armed bandits.
    Input: Exploration coefficient \(\rho \in \mathbb{R}_{+}\).
    At \(t=1, \ldots, K\), play each arm \(i \in\{1, \ldots, K\}\) once.
    for \(t \in\{K+1, K+2, \ldots\}\) do
        Play arm \(\pi_{t} \in \arg \max _{i \in\{1, \ldots, K\}}\left(\bar{X}_{i}(t-1)+\sqrt{\frac{\rho \log (t-1)}{N_{i}(t-1)}}\right)\).
```

Theorem 2 (Sampling rate of optimal arms in $K$-MAB under UCB) Fix $K \in \mathbb{N}$, and consider a $K$-armed model with arms indexed by $[K]:=\{1, \ldots, K\}$. Let $I \subseteq[K]$ be the set of optimal arms, i.e., arms with mean $\max _{i \in[K]} \mu_{i}$. If $\mathcal{I} \neq[K]$, define $\Delta_{\min }:=\max _{i \in[K]} \mu_{i}-\max _{i \in[K] \backslash I} \mu_{i}$. Then, there exists a finite $\rho_{0}>1$ that depends only on $|\mathcal{I}|$, such that the following results hold for any arm $i \in \mathcal{I}$ as $n \rightarrow \infty$ under Algorithm 2 initialized with $\rho \geqslant \rho_{0}$ :
(I) If $\mathcal{I}=[K]$, then

$$
\frac{N_{i}(n)}{n} \xrightarrow{p} \frac{1}{K} .
$$

(II) If $\mathcal{I} \neq[K]$ and optimal arms are "well-separated," i.e., $\Delta_{\min }=\omega\left(\sqrt{\frac{\log n}{n}}\right)$, then

$$
\frac{N_{i}(n)}{n} \xrightarrow{p} \frac{1}{|\mathcal{I}|}
$$

Discussion. The main observation here is that if the set of optimal arms is "sufficiently separated" from the sub-optimal arms, then classical UCB policies eventually allocate the sampling effort over the set of optimal arms uniformly, in probability. This is a desirable property to have from a fairness standpoint, and also markedly different from the instability and imbalance results for Thompson Sampling discussed earlier in Theorem 3. We remark that the $\rho \geqslant \rho_{0}$ condition is only necessary for tractability of the proof, and conjecture the result to hold, in fact, for any $\rho>1$, akin to the result for the two-armed setting (Theorem 1). We also conjecture analogous results for "small gap" and "moderate gap" regimes, in the spirit of Theorem 1; proofs, however, can be unwieldy in the general $K$-armed setting. The detailed proof of Theorem 2 is provided in §A. 9 .

What about Thompson Sampling? Results such as those discussed above for other popular adaptive algorithms like Thompson Sampling are only arable in "well-separated" instances where $N_{i^{*}}(n) / n \xrightarrow{p} 1$ as $n \rightarrow \infty$ follows as a trivial consequence of its $O(\sqrt{n})$ minimax regret bound. ${ }^{4}$ For smaller gaps, theoretical understanding of the distribution of arm-pulls under Thompson Sampling remains largely absent even for its most widely-studied variants. In this paper, we provide a first result in this direction: Theorem 3 formalizes a revealing observation for classical Thompson Sampling (Algorithm 3) in instances with zero gap, and elucidates its instability in view of the numerical evidence reported in Figure 1.1(b) and 1.1(c). This result also offers an explanation for the sharp contrast with the statistical behavior of canonical UCB (Algorithm 1) à la Theorem 1, also evident from Figure 1.1(a). In what follows, rewards are Bernoulli, and $S_{i}$ (respectively $F_{i}$ ) counts the number of successes/1's (respectively failures/0's) associated with arm $i \in\{1,2\}$.

```
Algorithm 3 Thompson Sampling for the two-armed Bernoulli bandit.
    Initialize: Number of successes (1's) and failures (0's) for each arm \(i \in\{1,2\},\left(S_{i}, F_{i}\right)=\)
    \((0,0)\).
    for \(t \in\{1,2, \ldots\}\) do
        Sample for each \(i \in\{1,2\}, \mathcal{T}_{i} \sim \operatorname{Beta}\left(S_{i}+1, F_{i}+1\right)\).
        Play arm \(\pi_{t} \in \arg \max _{i \in\{1,2\}} \mathcal{T}_{i}\) and observe reward \(r_{t} \in\{0,1\}\).
        Update success-failure counts: \(S_{\pi_{t}} \leftarrow S_{\pi_{t}}+r_{t}, F_{\pi_{t}} \leftarrow F_{\pi_{t}}+1-r_{t}\).
```

[^2]Theorem 3 (Incomplete learning under Thompson Sampling when $\Delta=0$ ) In a two-armed model where both arms yield rewards distributed as Bernoulli(q), the following holds under Algorithm 3 as $n \rightarrow \infty$ :
(I) If $q=0$ (i.e., all rewards are deterministic 0 ), then

$$
\frac{N_{1}(n)}{n} \Rightarrow \frac{1}{2} .
$$

(II) If $q=1$ (i.e., all rewards are deterministic 1 ), then

$$
\frac{N_{1}(n)}{n} \Rightarrow \text { Uniform on }[0,1] .
$$

Proof sketch. The proof of Theorem 3 relies on a careful application of two subtle properties of the Beta distribution (Fact 2 and Fact 3), stated and proved in Appendix A.2,A.10. For part (I), we invoke symmetry to deduce $\mathbb{E} N_{1}(n)=n / 2$, and use Fact 2 to show that the standard deviation of $N_{1}(n)$ is sub-linear in $n$, thus proving the stated assertion in (I). More elaborately, Fact 2 states for the reward configuration in (I) that the probability of playing arm 1 after it has already been played $n_{1}$ times, and arm $2 n_{2}$ times, equals $\left(n_{2}+1\right) /\left(n_{1}+n_{2}+2\right)$. This probability is smaller than $1 / 2$ if $n_{1}>n_{2}$, which provides an intuitive explanation for the fast convergence of $N_{1}(n) / n$ to $1 / 2$ observed in Figure 1.1(c). In fact, we conjecture that the result in (I) holds also with probability 1 based on the aforementioned "self-balancing" property. The conclusion in part (II) hinges on an application of Fact 3 to show the stronger result: $N_{1}(n)$ is uniformly distributed over $\{0,1, \ldots, n\}$ for any $n \in \mathbb{N}$. Contrary to Fact 2, Fact 3 states that quite the opposite is true for the reward configuration in (II): the probability of playing arm 1 after it has already been played $n_{1}$ times, and arm $2 n_{2}$ times, equals $\left(n_{1}+1\right) /\left(n_{1}+n_{2}+2\right)$, which is greater than $1 / 2$ when $n_{1}>n_{2}$. That is, the posterior distributions evolve in such a way that the algorithm is "deceived" into incorrectly believing one of the arms (arm 2 in this case) to be inferior. This leads to large sojourn times between successive visitations of arm 2 on such a sample-path, thereby resulting in a perpetual
"imbalance" in the sample-counts. This provides an intuitive explanation for the non-degeneracy observed in Figure 1.1(b) and 1.2(a), which additionally, also indicates that such behavior, in fact, persists also for general (non-deterministic) reward distributions, as well as under the Gaussian prior-based version of the algorithm. Full proof of Theorem 3 is provided in Appendix A.6.

More on "incomplete learning." The zero-gap setting is a special case of the "small gap" regime where canonical UCB guarantees a $(1 / 2,1 / 2)$ sample-split in probability (Theorem 1$)$. On the other hand, Theorem 3 suggests that second order factors such as the mean signal strength (magnitude of the mean reward) could significantly affect the nature of the resulting sample-split under Thompson Sampling. Note that even though the result only presupposes deterministic 0/1 rewards, the aforementioned claim is, in fact, borne out by the numerical evidence in Figure 1.1(b) and 1.1(c). The sampling distribution seemingly flattens rapidly from the Dirac measure at $1 / 2$ to the Uniform distribution on $[0,1]$ as the mean rewards move away from 0 . This uncertainty in the limiting sampling behavior has non-trivial implications for a variety of application areas of such learning algorithms. For instance, a Uniform distribution of arm-sampling rates on [0, 1] indicates that the sample-split could be arbitrarily imbalanced along a sample-path, despite, as in the setting of Theorem 3, the two arms being statistically identical; this phenomenon is typically referred to as "incomplete learning" in literature and has origins in [27, 28]. Non-degeneracy in the limiting distribution is also observable numerically up to diffusion-scale gaps of $O(1 / \sqrt{n})$ under other versions of Thompson Sampling (see [19] for examples); our focus on the more extreme zero-gap setting simplifies the illustration of these effects.

A brief survey of Thompson Sampling. While extant literature does not provide any explicit result for Thompson Sampling characterizing its arm-sampling behavior in instances with "small" and "moderate" gaps, there has been recent work on its analysis in the $\Delta_{n} \asymp 1 / \sqrt{n}$ regime under what is known as the diffusion approximation lens (see [19, 29]). Cited works study Thompson Sampling primarily under the assumption that the prior variance associated with the mean reward of any arm vanishes in the horizon of play at an "appropriate" rate; the non-vanishing variance set-
ting is amenable to analysis only under triangulation limits in general. ${ }^{5}$ Such a scaling, however, is not ideal from a regret standpoint and indeed, the versions of Thompson Sampling optimized for regret performance use fixed (non-vanishing) prior variances, e.g., Algorithm 3 and its Gaussian prior-based counterpart (see [7]). On a high level, [19, 29] establish that as $n \rightarrow \infty$, the pre-limit $\left(N_{i}(n t) / n\right)_{t \in[0,1]}$ under Thompson Sampling converges weakly to a "diffusion-limit" stochastic process on $t \in[0,1]$. Recall from earlier discussion that $\Delta_{n} \asymp 1 / \sqrt{n}$ is covered under the "small gap" regime; consequently, it follows from Theorem 1 that the analogous limit for UCB is, in fact, the deterministic process $t / 2$. In sharp contrast, the diffusion-limit process under Thompson Sampling may at best be characterizable only as a solution (possibly non-unique) to an appropriate stochastic differential equation or ordinary differential equation driven by a suitably (random) time-changed Brownian motion. Consequently, the diffusion limit under Thompson Sampling is more difficult to interpret vis-à-vis UCB (see Theorem 5), and it is much harder to obtain lucid insights as to the nature of the distribution of $N_{i}(n) / n$ as $n \rightarrow \infty$. The asymptotic distribution of $N_{i}(n) / n$ under Thompson Sampling is also investigated in [30], albeit in a significantly different setting. Cited paper considers the Bayesian setting where a prior distribution exists over problem instances, and the Thompson Sampling algorithm is "well-specified," i.e., information about said prior is baked into the algorithm. Specifically, a sample path of the algorithm in their model involves a random problem instance from the instance-space. In contrast, the derivation of the asymptotic distribution of arm-pulls in our work is for specific (fixed) problem instances, viz., reward configurations (I) and (II) described in Theorem 3, and under the classical version of Thompson Sampling (Algorithm 3).

### 1.2.2 Beyond arm-sampling rates

This part of the paper is dedicated to a more fine-grained analysis of the "stochastic" regret of UCB (defined in (1.1) in §1.1). Results are largely facilitated by insights on the sampling behavior of UCB in instances with "small" gaps, attributable to Theorem 1; however, we believe they are

[^3]of interest in their own right. We commence with an application of Theorem 1 which provides the first complete characterization of the worst-case (minimax) performance of canonical UCB. A full diffusion-limit characterization of the two-armed bandit problem under UCB is provided thereafter in Theorem 5.

Theorem 4 (Asymptotics of worst-case regret under UCB) In the "moderate gap" regime referenced in Theorem 1 where $\Delta_{n} \sim \sqrt{\theta \log n / n}$, the regret of the policy $\pi$ given by Algorithm 1 with $\rho>1$ satisfies

$$
\begin{equation*}
\frac{R_{n}^{\pi}}{\sqrt{n \log n}} \Rightarrow \sqrt{\theta}\left(1-\lambda_{\rho}^{*}(\theta)\right)=: h_{\rho}(\theta) \quad \text { as } n \rightarrow \infty \tag{1.3}
\end{equation*}
$$

where $\lambda_{\rho}^{*}(\theta)$ is the (unique) solution to (1.2).

To the best of our knowledge, this is the first algorithm-specific result (sharp asymptotic) that is distinct from the general $\Omega(\sqrt{n})$ information-theoretic lower bound by a horizon-dependent factor. ${ }^{6}$

Discussion. A closed-form expression for $\lambda_{\rho}^{*}(\theta)$ and $h_{\rho}(\theta)$ is provided in Appendix A.1. The behavior of $h_{\rho}(\theta)$ is illustrated below in Figure 1.3. For a fixed $\rho$, the function $h_{\rho}(\theta)$ is numerically observed to be uni-modal in $\theta$ and admit a global maximum at a unique $\theta_{\rho}^{*}:=\arg \sup _{\theta \geqslant 0} h_{\rho}(\theta)$, bounded away from 0 . Theorem 4 establishes that the worst-case (instance-independent) regret admits the sharp asymptotic $R_{n}^{\pi} \sim h_{\rho}\left(\theta_{\rho}^{*}\right) \sqrt{n \log n}$. In standard bandit parlance, this substantiates that the $O(\sqrt{n \log n})$ worst-case (minimax) performance guarantee of canonical UCB cannot be improved in terms of its horizon-dependence. In addition, the result also specifies the precise asymptotic constants achievable in the worst-case setting. This can alternately be viewed as a direct approach to proving the $O(\sqrt{n \log n})$ performance bound for UCB vis-à-vis conventional minimax analyses such as those provided in [2].

Proof sketch. On a high level, note that when $\Delta=\sqrt{(\theta \log n) / n}$, it follows from Theorem 1

[^4]

Figure 1.3: $\boldsymbol{h}_{\boldsymbol{\rho}}(\boldsymbol{\theta})$ vs. $\boldsymbol{\theta}$ for different values of the exploration coefficient $\rho$ in Algorithm 1. The graphs exhibit a unique global maximizer $\theta_{\rho}^{*}$ for each $\rho$. The ordered pairs $\left(\theta_{\rho}^{*}, h_{\rho}\left(\theta_{\rho}^{*}\right)\right)$ for $\rho \in\{1.1,2,3,4\}$ are $(1.9,0.24),(3.5,0.32),(5.3,0.39),(7,0.45)$.
that $\mathbb{E} R_{n}^{\pi}=\sqrt{(\theta \log n) / n \mathbb{E}}\left[n-N_{i^{*}}(n)\right] \sim h_{\rho}(\theta) \sqrt{n \log n}$ (convergence in probability implies that in mean since $\left.\left|N_{i^{*}}(n) / n\right| \leqslant 1\right)$. That $R_{n}^{\pi}$ also admits the same sharp asymptotic can be shown via a finer analysis. In other regimes of $\Delta$, viz., "small" and "large" gaps, we already know that $R_{n}^{\pi}=o_{p}(\sqrt{n \log n})$. This is obvious for "small" $\Delta$ since $\mathbb{E} R_{n}^{\pi} \leqslant \Delta n=o(\sqrt{n \log n})$, while for "large" $\Delta$, we use $\mathbb{E} R_{n}^{\pi} \leqslant C \rho((\log n) / \Delta+1 /(\rho-1))$ for some absolute constant $C$ (given $\rho>1$, $\Delta \leqslant 1$ ) [4], followed by Markov's inequality. Thus, the constant $\sup _{\theta \geqslant 0} h_{\rho}(\theta)$ obtained in the "moderate" gap regime must correspond to the worst-case performance of the algorithm.

Towards diffusion asymptotics. Diffusion scaling is a classical stochastic analysis tecnnique widely used in the mathematics and operations research literature, see, e.g., steady-state analyses of queuing systems in [22], and a recent application to certain sequential testing problems in [31]. Under this lens, time is accelerated linearly in $n$, space contracted by a factor of $\sqrt{n}$, and a sequence of systems indexed by $n$ is considered. In our problem, the $n^{\text {th }}$ such system refers to an instance of the two-armed bandit with: $n$ as the horizon of play; a gap that vanishes in the horizon as $\Delta_{n}=c / \sqrt{n}$ for some fixed $c$; and fixed reward variances given by $\sigma_{1}^{2}, \sigma_{2}^{2}$. This is a natural scaling for MAB experiments in that it "preserves" the hardness of the learning problem as $n$ sweeps over the sequence of systems. Recall also from previous discussion that the "hardest" information-theoretic instances have a $\Theta(1 / \sqrt{n})$ gap; in short, the diffusion limit is an appropriate asymptotic lens for observing interesting process-level behavior in the MAB problem. However, despite the aforementioned reasons, the diffusion limit behavior of bandit algorithms
remains poorly understood and largely unexplored. A recent foray was made in [19, 29], however, deterministic algorithms like UCB remain outside the ambit of such analysis on account of discontinuous arm-sampling probabilities. Theorem 5 provides a complete characterization of this limit for the celebrated UCB1 policy [10].

Theorem 5 (Diffusion asymptotics for canonical UCB) Suppose that the mean reward of arm $i \in$ $\{1,2\}$ is given by $\mu_{i}^{(n)}=\mu+\theta_{i} / \sqrt{n}$, where $n$ is the horizon of play and $\mu, \theta_{1}, \theta_{2} \geqslant 0$ are fixed constants, and reward variances are $\sigma_{1}^{2}, \sigma_{2}^{2}$. Define $\Delta_{0}:=\left|\theta_{1}-\theta_{2}\right|$. Denote the cumulative reward earned from arm $i$ until time $m$ by $S_{i, m}:=\sum_{j=1}^{N_{i}(m)} X_{i, j}$, and let $\tilde{S}_{i, m}:=S_{i, m}-\mu N_{i}(m)$. Then, the following process-level convergences hold under the policy $\pi$ given by Algorithm 1 with $\rho>1$ :

$$
\begin{aligned}
& \text { (I) }\left(\frac{\tilde{S}_{1,\lfloor n t\rfloor}}{\sqrt{n}}, \frac{\tilde{S}_{2,\lfloor n t\rfloor}}{\sqrt{n}}\right) \Rightarrow\left(\frac{\theta_{1} t}{2}+\frac{\sigma_{1}}{\sqrt{2}} B_{1}(t), \frac{\theta_{2} t}{2}+\frac{\sigma_{2}}{\sqrt{2}} B_{2}(t)\right), \\
& \text { (II) } \frac{R_{\lfloor n t\rfloor}^{\pi}}{\sqrt{n}} \Rightarrow \frac{\Delta_{0} t}{2}+\sqrt{\frac{\sigma_{1}^{2}+\sigma_{2}^{2}}{2}} \tilde{B}(t),
\end{aligned}
$$

where the process-level convergence is over $t \in[0,1]$, and $B_{1}(t)$ and $B_{2}(t)$ are independent standard Brownian motions in $\mathbb{R}$, and $\tilde{B}(t):=-\sqrt{\frac{\sigma_{1}^{2}}{\sigma_{1}^{2}+\sigma_{2}^{2}}} B_{1}(t)-\sqrt{\frac{\sigma_{2}^{2}}{\sigma_{1}^{2}+\sigma_{2}^{2}}} B_{2}(t)$.

Proof sketch. Note that if the arms are played $\lfloor n / 2\rfloor$ times each independently over the horizon of play $n$ (resulting in $N_{i}(n)=\lfloor n / 2\rfloor, i \in\{1,2\}$ ), part (I) of the stated assertion would immediately follow from Donsker's Theorem. However, since the sequence of plays, and hence also the eventual allocation $\left(N_{1}(n), N_{2}(n)\right)$, is determined adaptively by the policy, the aforementioned convergence may no longer be true. Here, the result hinges crucially on the observation from Theorem 1 that $N_{i}(n) / n \xrightarrow{p} 1 / 2$ under UCB when $\Delta_{n} \asymp 1 / \sqrt{n}$ (diffusion-scaled gaps are covered under the "small gap" regime). This observation facilitates a standard "random time-change" argument $t \leftarrow N_{i}(\lfloor n t\rfloor) / n, i \in\{1,2\}$, which followed upon by an application of Donsker's Theorem, leads to the assertion in (I). This has the profound implication that for diffusion-scaled gaps, a two-armed bandit under UCB is, in fact, well-approximated by a classical system with independent samples (sample-interdependence due to the adaptive nature of the policy is washed away in the limit). The
conclusion in (II) follows after an application of the Continuous Mapping Theorem to (I).
Discussion. An immediate observation following Theorem 5 is that the normalized regret $R_{n}^{\pi} / \sqrt{n}$ is asymptotically Normal with mean $\Delta_{0} / 2$ and variance $\left(\sigma_{1}^{2}+\sigma_{2}^{2}\right) / 2$ under UCB. Apart from aiding in obvious inferential tasks like construction of (asymptotically valid) confidence intervals (see, e.g., the binary hypothesis testing example referenced in Figure 1.2(c)), etc., such information provides new insights as to the problem's minimax complexity as well. This is because $\Delta_{n} \asymp 1 / \sqrt{n}$ is known to be the information-theoretic "worst-case" for the problem; the smallest achievable regret in this regime must asymptotically be dominated by that under UCB, i.e., $\Delta_{0} \sqrt{n} / 2$. It is also noteworthy that while the diffusion limit in Theorem 5 does not itself depend on the exploration coefficient $\rho$, the rate at which the system converges to said limit indeed depends on $\rho$. Theorem 5 will continue to hold only as long as $\rho=\omega((\log \log n) / \log n)$; for smaller $\rho$, the convergence of $N_{i}(n) / n$ to $1 / 2$ may no longer be true (refer to the proof in Appendix A.4).


Figure 1.4: Diffusion ensemble of UCB: x-axis is normalized time $t \in[0,1]$; y-axis is $\frac{R_{[n t]}^{\pi}}{\sqrt{n}}$ where $n=10,000$ rounds, and $\pi$ is Algorithm 1 with $\rho=2$.

### 1.3 Numerical experiments

This section provides additional numerical experiments illustrating the various "small gap" phenomena studied in this chapter. We consider the following three problem instances for experi-
ments:

- Instance 1: A 2-armed Bernoulli bandit with means $\mu_{1}=0.55$ and $\mu_{2}=0.5$.
- Instance 2: A 2-armed Bernoulli bandit with means $\mu_{1}=0.55$ and $\mu_{2}=0.51$.
- Instance 3: A 2-armed Bernoulli bandit with means $\mu_{1}=\mu_{2}=0.5$ each.

Each of these instances is played by Algorithm 1 with $\rho=2$ (same as UCB1 in [10]) and Algorithm 3 (Thompson Sampling for Bernoulli bandits, same as Algorithm 1 in [13]) separately. The horizon of play is limited to $T=10,000$ rounds, and we simulate $\boldsymbol{\mathcal { N }}=10,000$ independent sample-paths under each algorithm.

Performance metrics. We consider the following characteristics to highlight differences between the two algorithms.

- Distribution of regret: We plot the empirical distribution of pathwise regret after $T$ rounds, given by $R_{T}^{\pi}=\sum_{t=1}^{T}\left[\max \left(\mu_{1}, \mu_{2}\right)-X_{\pi_{t}, N_{\pi_{t}}(t)}\right]$, by simulating $\boldsymbol{\aleph}=10,000$ independent trajectories of each algorithm. We also show the $99 \%$ value-at-risk (VAR) and conditional-value-at-risk (CVAR) on these plots to highlight differences between the tail statistics of regret under the two algorithms. Refer to Figures 1.6, 1.7, 1.11, and 1.12.
- Arm-allocation statistics: We plot the empirical distribution of the fraction of time spent on arm 1 after $T$ rounds of play, given by $N_{1}(T) / T$, by simulating $\boldsymbol{\aleph}=10,000$ independent trajectories of each algorithm. Refer to Figures 1.8, 1.9, 1.13, and 1.14. We also plot the temporal evolution of $N_{1}(t) / t$ over $t \in\{1, \ldots, T\}$ (on a fixed sample-path) in Figures 1.5 and 1.10. Aforementioned figures also indicate the arm pulled at each $t$ along said sample-path.
- Evolution of posterior means and posterior distributions under Thompson Sampling when $\Delta=0$ (on a fixed sample-path): Refer to Figures 1.15 and 1.16.
- Diffusion-scaled regret: Figure 1.17 shows $\boldsymbol{\aleph}$ independent trajectories of the process $R_{\lfloor t T\rfloor}^{\pi} / \sqrt{T}$ over $t \in[0,1]$. Figure 1.18 shows the cross-sectional distribution at $t=1$.


Figure 1.5: Instance 1: x-axis is time; blue curve plots the evolution of $N_{1}(t) / t$ on a fixed samplepath; red dots indicate the arm pulled at time $t$.



Figure 1.6: Instance 1: Distribution of regret


Figure 1.7: Instance 1: Distribution of regret (right $1 \%$ tail)


Figure 1.8: Instance 1: Allocation statistics for arm 1


Figure 1.9: Instance 1: Allocation statistics for arm 1 (left 1\% tail)



Figure 1.10: Instance 2: x-axis is time; blue curve plots the evolution of $N_{1}(t) / t$ on a fixed samplepath; red dots indicate the arm pulled at time $t$.


Figure 1.11: Instance 2: Distribution of regret



Figure 1.12: Instance 2: Distribution of regret (right $1 \%$ tail)


Figure 1.13: Instance 2: Allocation statistics for arm 1



Figure 1.14: Instance 3: Allocation statistics for arm 1 (Histogram shows the empirical distribution of $N_{1}(T) / T$ after $T=10,000$ rounds plotted by simulating $\boldsymbol{\aleph}=10,000$ independent sample-paths)


Figure 1.15: Instance 3: Posterior evolution under Thompson Sampling (Sample-path\#1)


Figure 1.16: Instance 3: Posterior evolution under Thompson Sampling (Sample-path\#2)


Figure 1.17: Instance 3: Diffusion ensemble of UCB and Uniform Sampling (fair coin toss at each step); x-axis is normalized time $t \in[0,1]$; y-axis is $\frac{R_{[t T]}^{\pi}}{\sqrt{T}}$ where $T=10,000$ rounds.


Figure 1.18: Instance 2 (Cross-sectional statistics): Histograms show the empirical distribution of $\left(R_{T}^{\pi} / \sqrt{T}-c / 2\right) / \sqrt{\frac{\sigma_{1}^{2}+\sigma_{2}^{2}}{2}}$ where $T=10,000$ rounds, $c=1$, and $\sigma_{1}^{2}, \sigma_{2}^{2}$ are the reward variances.

### 1.4 Concluding remarks and open problems

This chapter summarizes the contributions in [32] for $K=2$. While a simple extension of results to the $K$-armed setting is provided in Theorem 2, the $K$-armed problem in full generality remains challenging to analyze. Under UCB, however, we do postulate a division of sampling effort within and across clusters of "similar" and "separated" arms (determined by their relative sub-optimality gaps) that is à la Theorem 1 . We expect that similar generalizations are possible also for Theorem 4 and Theorem 5. Under Thompson Sampling, on the other hand, things are
less obvious even in the two-armed setting. For example, in spite of compelling numerical evidence (refer, e.g., to Figure 1.1(b)) suggesting a plausible non-degenerate limiting distribution of arm-sampling rates when mean rewards are bounded away from 0 , the proof of Theorem 3 relies heavily on the rewards being deterministic and cannot be extended to the general stochastic case. In addition, similar results are conjectured also for the more widely used Gaussian prior-based version of the algorithm. These remain important open problems at the moment further progress on which would substantially improve our understanding of Thompson Sampling.

# Chapter 2: Multi-armed bandits with $K$ arm-types: A new countable-armed formulation 

### 2.1 Introduction

Background and motivation. In the classical multi-armed bandit (MAB) problem, the decision maker (DM) must play at each time instant $t \in\{1, \ldots, n\}$ one out of a set of $K \ll n$ possible alternatives ( $a k a \mathrm{arms}$ ), each characterized by a distribution of rewards. Oblivious to their statistical properties, the DM must play a sequence of $n$ arms so as to maximize her cumulative expected payoffs, an objective often converted to minimizing regret relative to an oracle with perfect ex ante knowledge of the best arm.

Since the seminal paper of [3] that laid the main theoretical foundations, there has been a plethora of work developing more advanced MAB models to encapsulate more realistic data-driven decision processes. These include formulations with covariate or contextual information, choicemodels, budget constraints, non-stationary rewards, and metric space embeddings, among many others that utilize some structure in the arms, reward distributions, or physics of the problem (see [33, 9] for a comprehensive survey). In this paper, we are motivated by the choice overload phenomenon pervading modern MAB applications with a prohibitively large action space such as those encountered in online marketplaces, matching platforms and the likes.

Modeling choice overload. In several applications of MAB, it is quite common for the number of arms to be "large" to the extent that it may potentially exceed even the horizon of play, i.e. $K \geqslant n$. For example, the problem faced by recommendation systems in large retail platforms, such as Amazon, is characterized by a prohibitively large number of arms (products of certain type) and limited "display space," creating a very challenging combinatorial problem (see, e.g., [34]). Naturally, the canonical MAB model is ill-suited for the study of such settings. Among problems
of this nature, a simple yet illuminating abstraction is one where an infinite population of arms is partitioned into $K$ different arm-types, each characterized uniquely by some reward statistic (e.g., the mean), and the fraction of each arm-type in the population of arms (aka the arm-reservoir) remains fixed over the horizon of play. The motivation to study such settings stems from several contemporary applications, e.g., in a prototypical task-matching problem arising in the online gig economy: the platform must choose upon each task arrival, one agent from a large pool of available agents characterized by unknown (or only partially known) skill proficiencies. Such settings arise naturally in populations endowed with latent low-dimensional representations, i.e., an agent can only belong to one of finitely many possible types, each characterized distinctly by some attribute. Market segmentation based on types is central also to the operations research literature, see, e.g., [35, 36, 37], etc., for examples involving analyses of online service and recommendation systems, among several other areas.

The countable-armed bandit problem. We provide an abstraction of the aforementioned decision problem as a bandit with countably many arms, each queried from an infinite population of arms (henceforth referred to as the arm-reservoir). There are $K$ possible arm-types in the reservoir given by $\mathcal{K}:=\{1, \ldots, K\}$, where $K$ is known a priori. Positing ex ante knowledge of $K$ is not unreasonable since it is routine for platforms to run pilot experiments during initial rounds to gather information on key primitives such as the size and stability of clusters, if any exist in the population. One can therefore safely assume in settings where such clusters strongly exist that the number $K$ of possible types is accurately estimated. The probability vector $\boldsymbol{\alpha}=\left\{\alpha_{i}: i \in \mathcal{K}\right\}$ denotes their corresponding fraction, i.e., relative prevalence in the reservoir, which is unknown.

Intuitively, the statistical complexity of regret minimization in the simplest formulation of the countable-armed bandit (CAB) with $K=2$ arm-types is governed by three principal primitives: (i) the sub-optimality gap $\underline{\Delta}:=\mu_{1}-\mu_{2}$ between the mean rewards $\mu_{1}>\mu_{2}$ of the optimal and inferior arm-types; (ii) the probability $\alpha_{1}$ of sampling from the population an arm of the optimal type; and (iii) the horizon of play $n$. Absent knowledge of $\left(\underline{\Delta}, \alpha_{1}\right)$, exploration is challenging owing to the "large" number of arms. In particular, in contrast with the classical two-armed bandit, in the CAB
problem, any finite selection of arms may only contain the mean $\mu_{2}$. Consequently, any algorithm limited to such a consideration set will suffer a linear regret. Absence of information on $\alpha_{1}$ further exacerbates the challenges in the study of the CAB problem. Specifically, how many arms must one query from the reservoir to collect at least one optimal arm is difficult to answer absent $\alpha_{1}$.

Contributions. There has been limited technical development in this area and the literature remains sparse. In this work, we resolve several foundational questions pertaining to the complexity of the countable-armed setting and provide a comprehensive understanding of various other aspects thereof. Our theoretical contributions can be projected along the following axes:
(i) Complexity of regret. We establish information-theoretic performance lower bounds that are order-wise tight (in the horizon $n$ ) in the instance-dependent setting (Theorem 6). In the instance-independent (minimax) setting, we answer affirmatively an open question on achievability of $\tilde{O}(\sqrt{n})$ regret when $K=2$ and show that this order is best achievable up to poly-logarithmic factors in $n$ (Theorem 9). In addition, we provide a uniform lower bound on achievable performance that is tight in $n$ and explicitly captures the scaling behavior w.r.t. the fraction $\alpha_{1}$ of optimal arms, and furthermore, has a novel non-information-theoretic proof based entirely on convex analysis (Theorem 7). Finally, we establish that the scaling of achievable regret w.r.t. $K$ must at least be $\Omega(K \log K)$ (Theorem 8); the $\log K$ factor reflects the increase in problem complexity vis-à-vis the classical $K$-armed problem.
(ii) Algorithm design. We design algorithms that achieve aforementioned regret guarantees relying only on knowledge of $K$ and are agnostic to information pertaining to the reward distributions as well as the frequency of occurrence of different types. Our design principles (Algorithms 4 and 5) are functionally distinct from extant work on finite-armed bandits which reflects in a fundamentally different scaling of regret (see Theorems 9 and 10). We also provide resolution to an outstanding design issue in extant literature for $K=2$ (see Algorithm 6 and Theorem 11).
(iii) Regret behavior and arm-type distribution. In contrast to some observations on related models in the literature that show a higher order than $\log n$ (instance-dependent) regret-behavior w.r.t. $\alpha$, we establish that when the learner has knowledge of $K$ but not of $\alpha$, one can still achieve
$O(\log n)$ regret where the dependence on $\alpha$ only manifests as an additive loss (Theorems 10 and 11).

Before proceeding with a formal description of our model, we provide a brief overview of related works below.

Extant literature on bandits with infinitely many arms. These problems involve an infinite population of arms and a fixed reservoir distribution over a (typically uncountable) set of arm-types; a common reward statistic (usually the mean) uniquely characterizes each arm-type. The infinite-armed bandit problem traces its roots to [38] where the problem was studied under Bernoulli rewards and a reservoir distribution of Bernoulli parameters that is Uniform on [0, 1]. Subsequent works have considered more general reward and reservoir distributions on $[0,1]$, see, e.g., [39, 40, 41, 42]. In terms of the statistical complexity of regret minimization, an uncountably rich set of arm-types is tantamount to the minimal achievable regret being polynomial in the horizon of play (see aforementioned references). In contrast, the recently studied models in [21, 43] that our work is most closely related to, are fundamentally simpler owing to a finite set of armtypes; this is central to the achievability of logarithmic (instance-dependent) regret in this class of problems. These two works are briefly discussed below.

The CAB problem first appeared in [21] together with an online adaptive policy achieving $O(\log n)$ regret when $K=2$. This policy is derived from UCB1 [10] and relies on certain newly discovered concentration and convergence properties thereof (see [32] for a detailed discussion of said properties). However, the analysis of this policy cannot be adapted to $K>2$ types; we will later provide an example with $K>2$ where the policy will likely run into issues that can be effectively mitigated by the algorithms proposed in this paper. There is also recent literature [43] on a related setting where the set of inferior arm-types may be arbitrary as long as it is $\underline{\Delta}$-separated from the optimal mean. However, ex ante knowledge of the proportion $\alpha_{1}$ of optimal arms is necessary to achieve logarithmic regret in this setting. This aspect distinguishes their setting from CAB and will be discussed at length later.

Lastly, a formulation of the countable-armed problem based on pure exploration, referred to as
the "heaviest coin identification problem," was studied in [44] for $K=2$ (see [45] for subsequent developments). In contrast, our problem is based on optimization of cumulative payoff (or regret); as a result, it shares little similarity with cited works.

Outline of this chapter. A formal description of the model is provided in $\S 2.2 ; \S 2.3$ discusses lower bounds on achievable performance. We propose our algorithms in $\S 2.4$ and state supporting theoretical guarantees. Numerical experiments are provided in $\S 2.5$. Proofs and auxiliary results are relegated to Appendix B.

### 2.2 Problem formulation

The set of arm-types is given by $\mathcal{K}=\{1, \ldots, K\}$, and the decision maker (DM) only knows the cardinality $K$ of $\mathcal{K}$. Each type $i \in \mathcal{K}$ is characterized by a unique mean reward $\mu_{i}$; the reservoir is thus characterized by the collection $\boldsymbol{\mu}:=\left\{\mu_{i}: i \in \mathcal{K}\right\}$ of possible mean rewards. Without loss of generality, we assume $\mu_{1}>\ldots>\mu_{K}$ and refer to type 1 as the optimal type (we may refer to the others as inferior types). Define $\bar{\Delta}:=\mu_{1}-\mu_{K}$ and $\underline{\Delta}:=\mu_{1}-\mu_{2}$ as the maximal and minimal sub-optimality gaps respectively, and $\delta:=\min _{1 \leqslant i<j \leqslant K}\left(\mu_{i}-\mu_{j}\right)$ as the minimal reward gap. Finally, $\boldsymbol{\alpha}:=\left(\alpha_{i}: i \in \mathcal{K}\right)$ denotes the vector of reservoir probabilities for each type (aka the reservoir distribution), coordinate-wise bounded away from 0 . These primitives will be important drivers of the statistical complexity of the regret minimization problem, as we shall later see. The horizon of play is $n$, and the DM must play one arm at each time $t \in\{1, \ldots, n\}$.

The set of arms that have been played up to and including time $t \in\{1,2, \ldots\}$ is denoted by $\mathcal{I}_{t}$ (and $\mathcal{I}_{0}:=\phi$ ). The set of actions available to the DM at time $t$ is given by $\mathcal{A}_{t}=\mathcal{I}_{t-1} \cup\left\{\right.$ new $\left._{t}\right\} ;$ the DM must either play an arm from $I_{t-1}$ at time $t$ or select the action "new ${ }_{t}$ " which corresponds to querying (and playing) a new arm from the reservoir. This new arm is optimal-typed with probability $\alpha_{1}$ and sub-optimal otherwise. The DM is oblivious to $\boldsymbol{\alpha}$, and furthermore precluded from observing the type of an arm upon query or play. A policy $\pi:=\left(\pi_{1}, \pi_{2}, \ldots\right)$ is an adaptive allocation rule that prescribes at time $t$ an action $\pi_{t}$ from $\mathcal{A}_{t}$ (possibly randomized). Each pull (or play) of an arm results in a stochastic reward. The sequence of rewards realized from the first
$k$ pulls of an arm labeled $i$ (henceforth called arm $i$ ) is denoted by $\left(X_{i, j}\right)_{1 \leqslant j \leqslant k}$; these are meanpreserving in time keeping the arm fixed, independent across arms and time, and take values in $[0,1]$. The natural filtration at time $t$, denoted by $\mathcal{F}_{t}$ and defined w.r.t. the sequence of rewards realized up to and including time $t$, is given by $\mathcal{F}_{t}:=\sigma\left\{\left(\pi_{s}\right)_{1 \leqslant s \leqslant t},\left\{\left(X_{i, j}\right)_{1 \leqslant j \leqslant N_{i}(t)}: i \in \mathcal{I}_{t}\right\}\right\}$ (with $\left.\mathcal{F}_{0}:=\phi\right)$, where $N_{i}(t)$ denotes the number of pulls of arm $i$ up to and including time $t$. The cumulative pseudo-regret of policy $\pi$ after $n$ plays is given by $R_{n}^{\pi}:=\sum_{t=1}^{n}\left(\mu_{1}-\mu_{\mathcal{T}}\left(\pi_{t}\right)\right.$, where $\mathcal{T}\left(\pi_{t}\right) \in \mathcal{K}$ denotes the type of the arm played by $\pi$ at time $t$; note that $R_{n}^{\pi}$ is a sample pathdependent quantity. The DM is interested in the classical problem of minimizing the expectation of the cumulative regret $\hat{R}_{n}^{\pi}:=\sum_{t=1}^{n}\left(\mu_{1}-X_{\pi_{t}, N_{\pi_{t}}(t)}\right)$, given by

$$
\begin{equation*}
\inf _{\pi \in \Pi} \mathbb{E} \hat{R}_{n}^{\pi}=\inf _{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{n}\left(\mu_{1}-X_{\pi_{t}, N_{\pi_{t}}(t)}\right)\right] \underset{(\dagger)}{=\inf _{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{n}\left(\mu_{1}-\mu_{\mathcal{T}}\left(\pi_{t}\right)\right)\right]=\inf _{\pi \in \Pi} \mathbb{E} R_{n}^{\pi}, ~, ~, ~} \tag{2.1}
\end{equation*}
$$

where $(\dagger)$ follows from the Tower property of expectation, the infimum is over policies satisfying the non-anticipation property $\pi_{t}: \mathcal{F}_{t-1} \rightarrow \mathcal{P}\left(\mathcal{A}_{t}\right)$ for $t \in\{1,2, \ldots\} ; \mathcal{P}\left(\mathcal{A}_{t}\right)$ denotes the probability simplex on $\mathcal{A}_{t}$. Accordingly, the expectation in (2.1) is w.r.t. all the possible sources of randomness in the problem (rewards, policy, and the arm-reservoir).

### 2.3 Lower bounds for natural policy classes

There are three fundamental primitives governing the complexity of achievable regret in this setting, viz., (i) the minimal sub-optimality gap $\underline{\Delta}$; (ii) the proportion $\alpha_{1}$ of the optimal arm-type in the reservoir; and (iii) the number of arm-types $K$. We next characterize lower bounds on achievable performance w.r.t. each of these primitives.

### 2.3.1 Achievable performance w.r.t. $\underline{\Delta}$ : Information-theoretic lower bounds

The statistical complexity of this problem setting is best illustrated via the paradigmatic case of $K=2$ and $\alpha_{1} \leqslant 1 / 2$. In this case, one anticipates the problem to be at least as hard as the classical two-armed bandit with a mean reward gap of $\underline{\Delta}$. Indeed, we establish this in Theorem 6 via
information-theoretic reductions adapted to handle a countable number of arms (proof is provided in Appendix B.1). In what follows, an instance $v$ of the problem refers to a collection of reward distributions with gap $\underline{\Delta}$; note that we are excluding the reservoir probabilities $\boldsymbol{\alpha}=\left(\alpha_{1}, \alpha_{2}\right)$ from the definition of an instance. Recall that $\alpha_{1}$ denotes the reservoir probability associated with the optimal mean reward, and $\alpha_{2}=1-\alpha_{1}$ is the probability of the inferior. We will overload the notation for expected cumulative regret slightly to emphasize its dependence on $v$ as well as $\boldsymbol{\alpha}$.

Definition 1 (Admissible policies when $\boldsymbol{K}=2$ ) A policy $\pi$ is deemed admissible if for any instance $v$, reservoir distributions $\boldsymbol{\alpha}=\left(\alpha_{1}, 1-\alpha_{1}\right), \boldsymbol{\alpha}^{\prime}=\left(\alpha_{1}^{\prime}, 1-\alpha_{1}^{\prime}\right)$, and horizon $n$, one has that $\mathbb{E} R_{n}^{\pi}\left(\nu, \alpha^{\prime}\right) \geqslant \mathbb{E} R_{n}^{\pi}(\nu, \boldsymbol{\alpha})$ whenever $\alpha_{1}^{\prime} \leqslant \alpha_{1}$. The set of such policies is denoted by $\Pi_{a d m}$.

We remark that the aforementioned definition is not restrictive in our problem setting since it is only natural that any reasonable policy should incur a larger cumulative regret (in expectation) in problems where the reservoir holds fewer optimal arms (in proportion).

Theorem 6 (Information-theoretic lower bounds when $K=2$ ) There exists an absolute constant $C>0$ such that the following holds under any $\pi \in \Pi_{a d m}$ and any $\boldsymbol{\alpha}$ with $\alpha_{1} \leqslant 1 / 2$ :

1. For any $\underline{\Delta}>0$, there exists a problem instance $v$ such that $\mathbb{E} R_{n}^{\pi}(v, \boldsymbol{\alpha}) \geqslant C \log n / \underline{\Delta}$ for large enough $n$, where the "large enough $n$ " depends exclusively on $\underline{\Delta}$.
2. For any $n \in \mathbb{N}$, there exists a problem instance $v$ such that $\mathbb{E} R_{n}^{\pi}(v, \alpha) \geqslant C \sqrt{n}$.

Distinction from classical MAB. Although the above result bears resemblance to classical information-theoretic lower bounds for finite-armed bandits, it is imperative to note that the setting has a fundamentally greater complexity that requires a more nuanced analysis vis-à-vis the finitearmed problem. Traditional proofs, as a result, cannot be tailored to our setting in a translational manner. To see this, note that when $\alpha_{1}$ is high, a query of the reservoir is very likely to return an arm of the optimal type; in the limit as $\alpha_{1} \rightarrow 1$, the problem becomes degenerate as all policies incur zero expected regret. Clearly, the problem cannot be harder than a two-armed bandit with gap $\underline{\Delta}$ uniformly over all values of $\alpha_{1}$. While we conjecture $\alpha_{1}<1$ to be a sufficient condition
for the existence of $\Omega(\log n / \underline{\Delta})$ instance-dependent and $\Omega(\sqrt{n})$ instance-independent (minimax) lower bounds, there are technical challenges due to probabilistic type associations over countably many arms. The restriction to $\alpha_{1} \leqslant 1 / 2$ and admissible policies (Definition 1) is then necessary for tractability of the proof and it remains unclear if this can be generalized further. Furthermore, when $K>2$, even defining an appropriate notion of admissibility à la Definition 1 is non-trivial and will likely involve dependencies on $\boldsymbol{\mu}$ in addition to $\boldsymbol{\alpha}$; pursuits in this direction are currently left to future work.

### 2.3.2 Achievable performance w.r.t. $\alpha_{1}$ : A uniform lower bound for front-loading

Though the bounds in Theorem 6 are tight in $n$ as we shall later see, they fail to provide any actionable insights w.r.t. $\alpha$. A natural question in the CAB setting is whether and in what manner does the presence of countably many arms affect achievable regret. In particular, how does the difficulty associated with finding optimal arms from the reservoir (and the dependence on the distribution $\boldsymbol{\alpha}$ ) come into play. Below, we propose a lower bound that explicitly captures this dependence, albeit with respect to a somewhat restricted policy class.

Theorem 7 ( $\alpha$-dependent lower bound for any $\boldsymbol{K} \geqslant 2$ ) Denote by $\tilde{\Pi}$ the class of policies under which the decision to query the arm-reservoir at any time $s \in\{1,2, \ldots\}$ is independent of $\mathcal{F}_{s-1}$. Then, for all problem instances $v$ with a minimal sub-optimality gap of at least $\underline{\Delta}>0$, one has

$$
\liminf _{n \rightarrow \infty} \inf _{\pi \in \tilde{\Pi}} \frac{\mathbb{E} R_{n}^{\pi}(v, \boldsymbol{\alpha})}{\log n} \geqslant \frac{\left(1-\alpha_{1}\right)^{2} \underline{\Delta}}{\alpha_{1}}
$$

Discussion. The proof is located in Appendix B.2. Several comments are in order. (i) The class $\tilde{\Pi}$, in particular, includes policies that front-load queries, i.e., query the reservoir upfront for a pre-specified number of arms and then deploy a regret minimizing routine on the queried set until the end of the playing horizon, see, e.g., the Sampling-UCB algorithm due to [43]. (ii) The cited paper also derives an information-theoretic $\Omega\left(\log n / \alpha_{1} \underline{\Delta}\right)$ lower bound based on a standard reduction to a hypothesis testing problem, although notably their setting is non-trivially distinct
from ours (this reflects starkly different sensitivities of achievable regret to $1 / \alpha_{1}$-scale, as we shall later see). Importantly though, akin to Theorem 6, their bound too, establishes the existence of an instance with logarithmic regret. On the other hand, the foremost noticeable aspect of Theorem 7 that differs from aforementioned results is that it provides a uniform lower bound over all instances that are at least $\underline{\Delta}$-separated in the mean reward, as opposed to merely establishing their existence. (iii) The presence of $\underline{\Delta}$ in the numerator (unlike traditional bounds where $\underline{\Delta}$ resides in the denominator) suggests that while this bound may be vacuous if $\underline{\Delta}$ is "small," it certainly becomes most relevant when $\underline{\Delta}$ is "well-separated." At the same time, it should be noted that Theorem 7 does not contradict the $O\left(\log n / \alpha_{1} \underline{\Delta}\right)$ upper bound (up to logarithmic factors in $\underline{\Delta}$ ) of Sampling-UCB; it merely provides a tool to separate regimes of $\underline{\Delta}$ where one bound captures the dominant effects vis-à-vis the other. (iv) A novelty of Theorem 7 lies in its proof, which differs from classical lower bound proofs in that it is based entirely on ideas from convex analysis as opposed to the information-theoretic and change-of-measure techniques hitherto used in the literature.

Remarks. (i) It is not impossible to avoid the $1 / \alpha_{1}$-scaling of the instance-dependent logarithmic regret. We will later show via an upper bound for one of our algorithms (ALG2(n)) that the $\alpha_{1}$-dependence can, in fact, be relegated to constant order terms (ALG2 ( $n$ ) queries arms adaptively based on sample-history and therefore does not belong to $\tilde{\Pi})$. Importantly, this will somewhat surprisingly establish that the instance-dependent logarithmic bound in Theorem 6 is optimal w.r.t. to its dependence on $\alpha_{1}$ (the scaling w.r.t. $\underline{\Delta}$, however, may not be best possible as forthcoming upper bounds suggest). (ii) Theorem 7 holds also for any arm-reservoir where the optimal type is at least $\underline{\Delta}$-separated from the rest, the nature of types (countable or uncountable) notwithstanding.

### 2.3.3 Achievable performance w.r.t. $K$ : The Bandit and the Coupon-collector

In the classical $K$-armed bandit problem, the (instance-dependent) regret scales linearly with the number of arms. We will next show that the $K$-typed countable-armed setting studied in this paper differs from its $K$-armed counterpart on account of a fundamentally distinct scaling of regret w.r.t. $K$. We will illustrate this by pivoting to a full information setting with one-sample learning,
i.e., a setting where the decision maker observes the mean reward of an arm immediately upon pulling it, but does not learn whether it is optimal. Under such a setting, the optimal policy $\pi^{*}$ for the $K$-armed problem will pull each of the $K$ arms once and subsequently commit the residual budget of play to the optimal arm, thus incurring a lifetime regret of $\mathbb{E} R_{\infty}^{\pi_{\infty}^{*}}=\sum_{i=2}^{K}\left(\mu_{1}-\mu_{i}\right)=\Theta(K)$. The optimal policy for the $K$-typed countable-armed setting will, analogously, keep querying new arms from the reservoir until it has collected one of each of the $K$ types, and will subsequently commit to the arm within said collection that has the best mean reward. In this case, regret will only accrue until the decision maker has pulled one arm of each type.

Theorem 8 (Regret scaling w.r.t. K) In the full information setting, the lifetime regret of any policy under reservoir distribution $\boldsymbol{\alpha}$ and mean reward vector $\boldsymbol{\mu}$ is at least $\sum_{i=2}^{K} \alpha_{i}\left(\mu_{1}-\mu_{i}\right) K \log K$.

If the reservoir distribution remains non-degenerate w.r.t. the optimal type, i.e., the fraction $\alpha_{1}$ of optimal arms in the reservoir remains bounded away from 1 as $K$ increases, it is ensured that $\sum_{i=2}^{K} \alpha_{i}\left(\mu_{1}-\mu_{i}\right)$ remains non-vanishing in $K$. Consequently, the lower bound in Theorem 8 grows as $\Omega(K \underline{\Delta} \log K)$.

This result establishes a fundamentally distinct scaling of regret w.r.t. $K$ in the countablearmed setting vis-à-vis the $K$-armed one (in the full information setting). When the true type of an arm is not immediately observable, one only expects the $\Omega(K \log K)$ scaling to exacerbate. In fact, when the learning horizon is $n$, we conjecture that regret grows at least as $\Omega(K \log K \log n)$, where the $\Omega(\cdot)$ is modulo gap-dependent constants. Characterizing the information-theoretic optimal rate, however, remains a challenging open problem. The proof of Theorem 8 is provided in Appendix B.3.

### 2.4 Gap and reservoir adaptive policies

As discussed, our goal here is to investigate regret achievable under $\boldsymbol{\mu}$-adaptive algorithms that are agnostic also to ex ante information on the distribution $\alpha$ of possible arm types. We propose two algorithms; $\operatorname{ALG1}(n)$ and $\operatorname{ALG2}(n)$, that are both predicated on ex ante knowledge of the horizon
of play $n$. §2.4.1 discusses the first of these, $\operatorname{ALG1}(n)$, which uses knowledge of $n$ to calibrate the duration of its exploration phases. ALG1 (n) serves as an insightful basal motif for algorithm design in that it satisfies the desiderata of an $O(\log n)$ instance-dependent regret for general $K \geqslant 2$ as well as an $\tilde{O}(\sqrt{n})$ instance-independent (minimax) regret when $K=2$; the latter property settling an open problem in the literature. However, its regret has a sub-optimal dependence on $\boldsymbol{\alpha}$. We leverage structural insights from the analysis of $\operatorname{ALG1}(n)$ to explore another design in ALG2( $n$ ) in §2.4.2, which determines its exploration phase lengths adaptively, as opposed to prespecifying them upfront. This new design guarantees the best possible dependence of regret on $\alpha$. Finally in §2.4.3, we discuss a fully sequential adaptive algorithm from extant literature for $K=2$, and propose a simple modification to rid it of a certain fragile assumption pertaining to ex ante knowledge of the support of reward distributions. We also provide new sharper bounds for the modified algorithm and discuss potential issues with its generalization to $K \geqslant 2$ vis-à-vis ALG1 ( $n$ ) and ALG2 (n).

### 2.4.1 Explore-then-commit with a pre-specified exploration schedule

In what follows (and all subsequent algorithms), a new arm is one that is freshly queried from the reservoir (an arm without a history of previous pulls). This arm belongs to type $i$ with probability $\alpha_{i}$ independent of the problem history thus far (collection of arms and types queried and the corresponding reward realizations until the current time).

Discussion of Algorithm 4. The foremost noticeable feature of this algorithm is the (nearly) exponentially increasing exploration schedule. Specifically, in the $k^{\text {th }}$ epoch, each of the $K$ arms in the consideration set is played $\left\lceil e^{2 \sqrt{k}} \log n\right\rceil$ times. It suffices to cap the size of the consideration set at $K$ since the decision maker is a priori aware of the existence of exactly $K$ arm-types in the reservoir. Upon completion of the $k^{\text {th }}$ epoch, the cumulative-difference-of-reward statistic for each of the $\binom{K}{2}$ arm-pairs is compared against a threshold of $2 e^{-\sqrt{k}} m$, where $2 e^{-\sqrt{k}}$ should be imagined as a proxy for a lower bound on the minimal reward gap $\delta$. If said statistic is small relative to the threshold for some pair, the pair is likely to contain arms of the same type (equal means), in which

```
Algorithm 4 ALG1(n) (Fixed design ETC)
    Input: Horizon of play \(n\).
    Set budget \(T=n\); set epoch counter \(k=1\).
    Initialize new epoch: Query \(K\) new arms; call it consideration set \(\mathcal{A}=\{1, \ldots, K\}\).
    Set exploration duration \(L=\left\lceil e^{2 \sqrt{k}} \log n\right\rceil\).
    \(m \leftarrow \min (L,\lfloor T / K\rfloor)\).
    Play each arm in \(\mathcal{A} m\) times; observe rewards \(\left\{\left(X_{1, j}, \ldots, X_{K, j}\right): j=1, \ldots, m\right\}\).
    Update budget: \(T \leftarrow T-K m\).
    if \(\exists a, b \in \mathcal{A}, a<b\) s.t. \(\left|\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)\right|<2 m e^{-\sqrt{k}}\) then
        Permanently discard \(\mathcal{A}\).
        \(k \leftarrow k+1\).
        Repeat from step (3).
    else
        Permanently commit to \(\operatorname{arm} a^{*} \in \arg \max _{a \in \mathcal{A}}\left\{\sum_{j=1}^{m} X_{a, j}\right\}\).
```

case, the algorithm discards the entire consideration set and ushers in a new epoch with a larger exploration phase. This is done to avoid the possibility of incurring linear regret should an optimal arm be missing from the consideration set (e.g., when all $K$ arms belong to type 2 ). On the other hand, if all arm-pairs are sufficiently separated, the algorithm simply commits permanently to the empirically best arm. The intuition behind the (nearly) exponential schedule is that as $k$ grows, $2 e^{-\sqrt{k}}$ will eventually provide a lower bound on $\delta$, and one may hope to achieve appropriate levels of error control using window sizes in the $k^{\text {th }}$ epoch. Full proof is provided in Appendix B.4.

Theorem 9 (Upper bound on the regret of $\operatorname{ALG1}(\mathbf{n})$ ) For a horizon of play $n \geqslant K$, the expected cumulative regret of the policy $\pi$ given by $\operatorname{ALGI}(n)$ is bounded as

$$
\mathbb{E} R_{n}^{\pi} \leqslant \frac{\tilde{C}_{\alpha} \bar{\Delta} \log n}{\delta^{2}} \log ^{2}\left(\frac{4}{\delta}\right)+2 K \bar{\Delta},
$$

where $\tilde{C}_{\boldsymbol{\alpha}}$ is some constant that depends only on $\boldsymbol{\alpha}$; an exact expression is provided in (B.11). In particular, $\tilde{C}_{\boldsymbol{\alpha}} \rightarrow \infty$ as $\prod_{i=1}^{K} \alpha_{i}$ approaches 0.

Discussion. The dependence on the minimal reward gap $\delta$ in Theorem 9 is not an artifact of our analysis but, in fact, reflective of the operating principle of the algorithm. ALG1 ( $n$ ) keeps querying new consideration sets of size $K$ until it determines with high enough confidence that the queried
arms are all distinct-typed; this is the genesis of $\delta$ in the upper bound. Importantly, equipped just with knowledge of $K$, it remains unclear if there exists an alternative sampling strategy that does not rely on assessing pairwise similarities between the queried arms, without necessitating any additional information on $\boldsymbol{\alpha}$. Furthermore, while $\operatorname{ALG1}(n)$ is evidently rate-optimal in $n$ (in the instance-dependent sense), the scaling of its upper bound w.r.t. $\alpha$ is far from optimal. In particular, the $\alpha$-dependent factor in the leading term is attributable to a naive pre-determined exploration schedule. This dependence can, in fact, be relegated to $O(1)$ terms using a more sophisticated policy that operates based on an adaptive determination of stopping and re-initialization times.

Remarks. (i) When $K=2$, the upper bound in Theorem 9 reduces to $\left(\tilde{C}_{\alpha} / \underline{\Delta}\right) \log ^{2}(4 / \underline{\Delta}) \log n+$ $4 \underline{\Delta}$, leading to a worst-case regret (w.r.t. $\underline{\Delta}$ ) of $\tilde{O}\left(\tilde{C}_{\alpha} \sqrt{n}\right)$, where the big-Oh only hides polylogarithmic factors in $n$. This settles an open question concerning the best achievable minimax regret in the countable-armed problem with two types (since previously known guarantees were polynomially bounded away from $\sqrt{n}$; see [21]). (ii) While specifying the exploration schedule, the choice of the exponent in $k$ can be fairly general as long as it is coercive and grows sufficiently fast but sub-linearly; the square-root function is chosen for technical convenience. Instead, if one were to use a linear function of $k$ in the exponent, the algorithm's performance would become fragile w.r.t. ex ante knowledge of $\boldsymbol{\alpha}$; an ill-calibrated $\operatorname{ALG1}(n)$ can potentially incur linear regret.

### 2.4.2 Explore-then-commit with an adaptive stopping time

Discussion of Algorithm 5. At any time, the algorithm computes two thresholds; $O(\sqrt{m \log m})$ and $O(\sqrt{m \log n})$ for the $\binom{K}{2}$ pairwise difference-of-reward statistics, $m$ being the per-arm sample count. If said statistic is dominated by the former threshold for some arm-pair, it is likely to contain arms of the same type (equal means). The explanation stems from the Law of the Iterated Logarithm (see [25], Theorem 8.5.2): a zero-drift length- $m$ random walk has its envelope bounded by $O(\sqrt{m \log \log m})$. In the aforementioned scenario, the algorithm discards the entire consideration set and ushers in a new one. This is done to avoid the possibility of incurring linear regret should an optimal arm be missing from the consideration set. In the other scenario that the difference-of-

```
Algorithm 5 ALG2(n) (ETC with adaptive stopping times)
    Input: Horizon of play \(n\).
    Set budget \(T=n\); Burn-in samples (per arm) \(s_{n}\).
    Initialize new epoch: Query \(K\) new arms; call it consideration set \(\mathcal{A}=\{1, \ldots, K\}\).
    Play each \(\operatorname{arm}\) in \(\mathcal{A}\) for \(s_{n}\) periods; observe rewards \(\left\{X_{a, j}: a \in \mathcal{A}, j=1, \ldots, s_{n}\right\}\).
    Set per-arm sample count \(m=s_{n}\).
    Update budget: \(T \leftarrow T-s_{n} K\).
    Generate \(\binom{K}{2}\) independent standard Gaussian random variables \(\left\{\mathcal{Z}_{a, b}: a, b \in \mathcal{A}, a<b\right\}\).
    while \(T \geqslant K\) do
        if \(\exists a, b \in \mathcal{A}, a<b\) s.t. \(\left|\mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)\right|<4 \sqrt{m \log m}\) then
            Permanently discard \(\mathcal{A}\) and repeat from step (3).
        else
            if \(\left|\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)\right| \geqslant 4 \sqrt{m \log n} \forall a, b \in \mathcal{A}, a<b\) then
                    Permanently commit to arm \(a^{*} \in \arg \max _{a \in \mathcal{A}}\left\{\sum_{j=1}^{m} X_{a, j}\right\}\).
                else
                    Play each arm in \(\mathcal{A}\) once; observe rewards \(\left\{X_{a, m+1}: a \in \mathcal{A}\right\}\).
                    \(m \leftarrow m+1\).
            \(T \leftarrow T-K\).
```

reward statistic dominates the larger threshold for all arm-pairs, the consideration set is likely to contain arms of distinct types (no two have equal means) and the algorithm simply commits to the empirically best arm. Lastly, if difference-of-reward lies between the two thresholds (signifying insufficient learning), the sample count for each arm is advanced by one, and the entire process repeats.

Reason for introducing zero-mean corruptions supported on $\mathbb{R}$. Centered Gaussian noise is added to the difference-of-reward statistic in step (9) of ALG2 $(n)$ to avoid the possibility of incurring linear regret should the support of the reward distributions be a "very small" subset of $[0,1]$. To illustrate this point, suppose that $K=2, s_{n}=1$, and the rewards associated with the two types are deterministic with $\underline{\Delta}<2 \sqrt{2 \log 2}$. Then, as soon as the algorithm queries a heterogeneous consideration set (one arm optimal and the other inferior) and the per-arm sample count reaches 2, the difference-of-reward statistic will satisfy $\left|\sum_{j=1}^{2}\left(X_{1, j}-X_{2, j}\right)\right|=2 \underline{\Delta}<4 \sqrt{2 \log 2}$, resulting in the consideration set getting discarded. On the other hand, if the consideration set is homogeneous (both arms simultaneously optimal or inferior), the algorithm will still re-initialize as soon as the per-arm count reaches 2 ; this is because $\left|\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right|=0$ identically in this case for any
$m \in \mathbb{N}$ while $4 \sqrt{m \log m}>0$ only for $m \geqslant 2$. This will force the algorithm to keep querying new arms from the reservoir at rate that is linear in time, which is tantamount to incurring linear regret in the horizon. The addition of centered Gaussian noise hedges against this risk by guaranteeing that the difference-of-reward process essentially has an infinite support at all times even when the reward distributions might be degenerate. This rids the regret performance of its fragility w.r.t. the support of reward distributions. The next proposition crystallizes this discussion; proof is provided in Appendix B.5.

Proposition 1 (Persistence of heterogeneous consideration sets) Suppose $\left\{X_{a, j}: j=1,2, \ldots\right\}$ is a collection of independent samples from an arm of type $a \in\{1, \ldots, K\}=: \mathcal{A}$. Further suppose that $\left\{\mathcal{Z}_{a, b}: a, b \in \mathcal{A}, a<b\right\}$ is a collection of $\binom{K}{2}$ independently generated standard Normal random variables. Then,

$$
\begin{equation*}
\mathbb{P}\left(\bigcap_{m \geqslant 1} \bigcap_{a, b \in \mathcal{A}, a<b}\left\{\left|\mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)\right| \geqslant 4 \sqrt{m \log m}\right\}\right)>\frac{\bar{\Phi}\left(f\left(T_{0}\right)\right)}{2}=: \beta_{\delta, K}>0, \tag{2.2}
\end{equation*}
$$

where $\bar{\Phi}(\cdot)$ is the right tail of the standard Normal CDF, and $T_{0}:=\max \left(\left[\left(64 / \delta^{2}\right) \log ^{2}\left(\frac{64}{\delta^{2}}\right)\right], \Lambda_{K}\right)$ with $\Lambda_{K}:=\inf \left\{p \in \mathbb{N}: \sum_{m=p}^{\infty} \frac{1}{m^{8}} \leqslant \frac{1}{2 K^{2}}\right\}$. Lastly, $f(x):=x+4 \sqrt{x \log x}$ for all $x \geqslant 1$.

Interpretation of $\boldsymbol{\beta}_{\delta, K}$. First of all, note that $\beta_{\delta, K}$ admits a closed-form characterization in terms of standard functions and satisfies $\beta_{\delta, K}>0$ for $\delta>0$ with $\lim _{\delta \rightarrow 0} \beta_{\delta, K}=0$. Secondly, $\beta_{\delta, K}$ depends exclusively on $\delta$ and $K$, and represents a lower bound on the probability that ALG2( $n$ ) will never discard a consideration set containing arms of distinct types. This meta-result will be key to the upper bound on the regret of $\operatorname{ALG} 2(n)$ stated next in Theorem 10.

Theorem 10 (Upper bound on the regret of ALG2(n)) For a horizon of play $n \geqslant K$ and perarm burn-in phase of $1 \leqslant s_{n} \leqslant n / K$ samples, the expected cumulative regret of the policy $\pi$ given by $\operatorname{ALG2(n)}$ is bounded as

$$
\mathbb{E} R_{n}^{\pi} \leqslant \frac{C K^{3} \bar{\Delta}}{\gamma\left(s_{n}\right)}\left(\frac{\log n}{\delta^{2}}+\frac{s_{n}}{K!\prod_{i=1}^{K} \alpha_{i}}\right),
$$

where $C$ is some absolute constant, and $\gamma\left(s_{n}\right)$ is as defined in (B.12). In particular, $\gamma(t)$ is monotone increasing in $t$ with $\gamma(t) \rightarrow 1$ as $t \rightarrow \infty$, and $\gamma(1)=\beta_{\delta, K}$, where the latter is as defined in (2.2).

Remarks. The dependence on $\delta$ in Theorem 10 is not incidental and has the same genesis as discussed in the context of Theorem 9. However, there is a prominent distinction from Theorem 9 in that the dependence on $\alpha$ is captured exclusively through the constant term (as opposed to the logarithmic term). This should be viewed in light of the lower bound in Theorem 7; by allowing for policies that query the arm-reservoir adaptively, one can potentially make the regret performance robust w.r.t. $\alpha$. Absence of $\alpha$ from the leading term also leads to the somewhat remarkable conclusion that the lower bound in Theorem 6 is optimal w.r.t. dependence on $\boldsymbol{\alpha}$. The proof is provided in Appendix B.6.

More on the inverse scaling w.r.t. $\gamma\left(s_{\boldsymbol{n}}\right)$. This multiplicative factor is likely a consequence of the countable nature of arms (as opposed to finite). When $K=2, \alpha_{1} \leqslant 1 / 2$, and the burnin phase $s_{n}$ has a fixed duration independent of $n$, the upper bound in Theorem 10 reduces to $O\left(\beta_{\underline{\Delta}, 2}^{-1}\left(\log n / \underline{\Delta}+\underline{\Delta} / \alpha_{1}\right)\right)$, where the big-Oh only hides absolute constants. Evidently, there is an inflation by $\beta_{\Delta, 2}^{-1}$ relative to the optimal $O(\log n / \underline{\Delta})$ rate achievable in the paradigmatic two-armed bandit with gap $\underline{\Delta}$. By setting $s_{n}$ as a coercive sub-logarithmic function of the horizon $n$ (e.g., $s_{n}=$ $\sqrt{\log n})$, one can shave off the $\beta_{\underline{\Delta, 2}}^{-1}$ factor to achieve $O(\log n / \underline{\Delta})$ regret. This establishes tightness of the instance-dependent lower bound in Theorem 6 when $K=2$. On the other hand, owing to the dependence of $\gamma\left(s_{n}\right)$ (and $\beta_{\underline{\Delta}, 2}$ ) on $\underline{\Delta}$, the worst-case (instance-independent) upper bound of ALG2 ( $n$ ) can be observed from Theorem 10 to be bounded away from $\Omega(\sqrt{n})$. However, recall that Theorem 9 already settles the issue of characterizing the optimal minimax rate when $K=2$ (up to logarithmic factors in $n$ ). Thus, we provide a complete characterization of the complexity of this problem when $K=2$, thereby answering all the open problems in [21]. For $K>2$, Theorem 10 guarantees an upper bound of $O\left(K^{3} \bar{\Delta} / \delta^{2} \log n\right)$ under a coercive sub-logarithmic burn-in phase. In this case, characterizing the optimal dependence on $\bar{\Delta}$ and $\delta$ remains an open problem. The scaling w.r.t. $K$, however, cannot be improved to $O(K)$ as suggested by the $\Omega(K \log K)$ lower bound in

Theorem 8. A full characterization of the complexity of the general setting with $K>2$ arm-types remains challenging and is left to future work.

### 2.4.3 Towards fully sequential adaptive strategies: Optimism in exploration

In this section, we revisit the UCB-based adaptive policy proposed in [21] for $K=2$. The policy is restated as ALG3 below after suitable modifications for reasons discussed next. The original policy (Algorithm 2 in cited paper) achieves an instance-dependent regret of $O(\log n)$ and additionally enjoys the benefit of being anytime in $n$. However, it suffers a major limitation through its dependence on ex ante knowledge of the support of reward distributions. In particular, the algorithm requires the reward distributions to have "full support" on [0, 1], e.g., only distributions such as Bernoulli $(\cdot)$, Uniform on [0, 1], Beta $(\cdot, \cdot)$, etc., are amenable to its performance guarantees; Uniform on $[0,0.5]$, on the other hand, is not. We identify a simple fix to this issue: Drawing inspiration from the design of ALG2 ( $n$ ), we propose adding centered Gaussian noise to the difference-of-reward statistic in step (6) of Algorithm 6 to essentially create an unbounded support. This rids the algorithm of its fragility while preserving $O(\log n)$ regret (see Theorem 11).

Remark. The original Algorithm 2 in cited paper uses a different threshold in step (6) of Algorithm 6; the choice of $4 \sqrt{m \log m}$ here aims to unify the technical presentation with $\operatorname{ALG} 2(n)$, and facilitate a transparent comparison between their upper bounds.

```
Algorithm 6 ALG3 (Nested UCB1 for \(\mathrm{K}=2\) types)
    Initialize new epoch (resets clock \(t \leftarrow 0\) ): Query two new arms; call it set \(\mathcal{A}=\{1,2\}\).
    Play each arm in \(\mathcal{A}\) once; observe rewards \(\left\{X_{a, 1}: a \in \mathcal{A}\right\}\).
    Minimum per-arm sample count \(m \leftarrow 1\).
    Generate a standard Gaussian random variable \(\mathcal{Z}\).
    for \(t \in\{3,4, \ldots\}\) do
        if \(\left|\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right|<4 \sqrt{m \log m}\) then
            Permanently discard \(\mathcal{A}\) and repeat from step (1).
        else
            Play arm \(a_{t} \in \arg \max _{a \in \mathcal{A}}\left(\frac{\sum_{j=1}^{N_{a}(t-1)} X_{a, j}}{N_{a}(t-1)}+\sqrt{\frac{2 \log (t-1)}{N_{a}(t-1)}}\right)\).
            Observe reward \(X_{a_{t}, N_{a_{t}}(t)}\).
            if \(m<\min _{a \in \mathcal{A}} N_{a}(t)\) then
                \(m \leftarrow m+1\).
```

Discussion of Algorithm 6. Similar to ALG2 (n), ALG3 also has an episodic dynamic with exactly one pair of arms played per episode. The distinction, however, resides in the fact that ALG3 plays arms according to UCB1 in every episode as opposed to playing them equally often until committing to the empirically superior one. Secondly, unlike ALG2(n), ALG3 never "commits" to an arm (or a consideration set). The implication is that the algorithm will keep querying new consideration sets throughout the playing horizon; this property is at the core of its anytime nature. Despite these differences, the performance guarantees of the two algorithms are essentially identical when $K=2$, as the next result illustrates. The proof is provided in Appendix B.8.

Theorem 11 (Upper bound on the regret of ALG3 when $\boldsymbol{K}=\mathbf{2}$ ) The expected cumulative regret of the policy $\pi$ given by ALG3 after any number of pulls $n \geqslant 2$ is bounded as

$$
\mathbb{E} R_{n}^{\pi} \leqslant \frac{C}{\beta_{\underline{\Delta}, 2}}\left(\frac{\log n}{\underline{\Delta}}+\frac{\underline{\underline{\Delta}}}{\alpha_{1}}\right),
$$

where $\beta_{\underline{\Delta}, 2}$ is as defined in (2.2) with $\delta \leftarrow \underline{\Delta}$ and $K \leftarrow 2$, and $C$ is some absolute constant.

Remark. It is possible to shave off the $\beta_{\underline{\Delta}, 2}$ factor by introducing in ALG3 a horizon-dependent burn-in phase à la ALG2 (n). This may be achieved at the expense of ALG3's anytime property.

The performance stated in Theorem 11 together with its anytime property might appear to give an edge to ALG3 over ALG2(n). However, the former is theoretically disadvantaged in that its logarithmic upper bound is not currently amenable to extensions to the general $K$-typed setting. The adaption of ALG3 to the full generality of $K$ arm-types (see ALG4 below) does not currently admit a logarithmic upper bound due to reasons discussed subsequently.

```
Algorithm 7 ALG4 (Nested UCB1 for \(K\) types)
    Initialize new epoch (resets clock \(t \leftarrow 0\) ): Query \(K\) new arms; call it set \(\mathcal{A}=\{1, \ldots, K\}\).
    Play each arm in \(\mathcal{A}\) once; observe rewards \(\left\{X_{a, 1}: a \in \mathcal{A}\right\}\).
    Minimum per-arm sample count \(m \leftarrow 1\).
    Generate \(\binom{K}{2}\) independent standard Gaussian random variables \(\left\{\mathcal{Z}_{a, b}: a, b \in \mathcal{A}, a<b\right\}\).
    for \(t \in\{K+1, K+2, \ldots\}\) do
    if \(\exists a, b \in \mathcal{A}, a<b\) s.t. \(\left|\mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)\right|<4 \sqrt{m \log m}\) then
            Permanently discard \(\mathcal{A}\) and repeat from step (1).
        else
            Play arm \(a_{t} \in \arg \max _{a \in \mathcal{H}}\left(\frac{\sum_{j a=1}^{N_{a}(t-1)} X_{a, j}}{N_{a}(t-1)}+\sqrt{\frac{2 \log (t-1)}{N_{a}(t-1)}}\right)\).
            Observe reward \(X_{a_{t}, N_{a_{t}}(t)}\).
            if \(m<\min _{a \in \mathcal{A}} N_{a}(t)\) then
                \(m \leftarrow m+1\).
```

Theorem 12 (Upper bound on the regret of ALG4) The expected cumulative regret of the policy $\pi$ given by ALG4 after any number $n \geqslant 1$ of plays is bounded as

$$
\mathbb{E} R_{n}^{\pi} \leqslant \frac{C K}{\beta_{\delta, K}}\left(\frac{\log n}{\underline{\Delta}}+\bar{\Delta}\right)+o\left(\frac{\bar{\Delta} n}{\beta_{\delta, K} \prod_{i=1}^{K} \alpha_{i}}\right),
$$

where $C$ is some absolute constant, $\beta_{\delta, K}$ is as defined in (2.2), and the little-Oh is asymptotic in $n$ and only hides multiplicative factors in $K$.

Proof of Theorem 12 is provided in Appendix B.10.
Limitation of ALG4. The issue of a non-logarithmic upper bound traces its roots to the use of UCB1 as a subroutine. The concentration behavior of UCB1 leveraged towards the analysis of ALG3 when $K=2$ fails to hold when $K>2$, rendering proofs intractable. This is illustrated via a simple example with $K=3$ types discussed below.

Technical issues with generalizing ALG3 to $\boldsymbol{K}$ types. When $K=2$, there are only two possibilities for what a consideration set could be; arms can have means that are either (i) distinct,
or (ii) equal. In the former case, an optimal arm is guaranteed to exist in the consideration set and UCB1 will spend the bulk of its sampling effort on it, which is good for regret performance. In the latter scenario, since arms have equal means, UCB1 will split samples approximately equally between the two with high probability (see Theorem 4(i) in [21]); subsequently the consideration set will be discarded within a finite number of samples in expectation (see steps (6) and (7) of ALG3). Contrast this with an alternative setting with $K=3$ and mean rewards $\mu_{1}>\mu_{2}>\mu_{3}$. A natural generalization of ALG3 (see ALG4) will query consideration sets of size 3. Thus, a query can potentially return one arm with mean $\mu_{2}$ and two with mean $\mu_{3}$. Since an optimal arm (mean $\left.\mu_{1}\right)$ is missing, the algorithm will incur linear regret on this set; it is therefore imperative to discard it at the earliest. Unfortunately though, UCB1 will invest an overwhelming majority of its sampling effort in the "locally optimal" arm (mean $\mu_{2}$ ) and allocate logarithmically fewer samples among the other two. This logarithmic rate of sampling arms with mean $\mu_{3}$ is proof-inhibiting (vis-à-vis the $K=2$ case where the rate is linear as previously discussed), making it difficult to theoretically answer if ALG 4 might still be able to discard the arms within, say, logarithmically many pulls of the horizon. This is an open research question and at the moment, an $O(\log n)$ bound exists only for $K=2$; we could only establish asymptotic-optimality ( $o(n)$ regret) when $K>2$ (see Theorem 12). Among other things, identifying the optimal (instance-dependent) scaling factors w.r.t. ( $\boldsymbol{\mu}, \boldsymbol{\alpha}$ ) and the optimal order of minimax regret when $K>2$ remain open problems. Numerical experiments validating the bounds studied in this work are discussed in the next section.

### 2.5 Numerical experiments

We evaluate the empirical performance of our algorithms for $K=2$ and $K=3$ on synthetic data.

Experiments. In what follows, the graphs show the performance of different algorithms simulated on synthetic data. The horizon is capped at $n=10^{5}$ for $K=2$ and at $n=10^{4}$ for $K=3$. Each regret trajectory is averaged over at least 100 independent experiments (sample-paths). The shaded regions indicate standard $95 \%$ confidence intervals. For horizon-dependent algorithms, re-
gret is plotted for discrete values of the horizon $n$ indicated by "*" and interpolated; for anytime algorithms, regret accrued until each $t \in\{1, \ldots, n\}$ is plotted.

Baseline policies. We will benchmark the performance of our algorithms against two policies: (i) Sampling-UCB [43], and (ii) ETC-D(2) [21]. The former is a UCB-styled policy based on front-loading exploration of new arms (Theorem 7 thus applies to this policy). It is, however, noteworthy that Sampling-UCB is predicated on ex ante knowledge of (a lower bound on) the probability $\alpha_{1}$ of sampling an optimal arm from the reservoir; we reemphasize that this is not the setting of interest in our paper. Furthermore, its regret scales as $\tilde{O}\left(\log n /\left(\alpha_{1} \underline{\Delta}\right)\right)$ (up to polylogarithmic factors in $1 / \underline{\Delta}$ ), which is inferior in terms of its dependence on $\alpha_{1}$ relative to ALG2 ( $n$ ) and ALG3 (see Theorem 10 and 11 respectively). There exist other algorithms as well (see, e.g., [46, 47]) developed for formulations with prohibitively large number of arms. However, these are either sensitive to certain parametric assumptions on the probability of sampling an optimal arm, or focus on a different notion of regret altogether; both directions remain outside the ambit of our setting.

The second policy ETC- $\infty$ (2) is a non-adaptive explore-then-commit-styled algorithm for reservoirs with $K=2$ types; this policy requires ex ante knowledge of a lower bound on the difference between the two mean rewards. Although ETC $-\infty$ (2) was originally proposed only for $K=2$, it is easily generalizable and we present in Algorithm 8 below a version (ETC $-\infty(K)$ ) that is adapted to $K$ types.

```
Algorithm 8 ETC- \(\infty\) (K)
    Input: (i) Horizon of play \(n\), (ii) A lower bound \(\underline{\delta} \in(0, \delta]\) on the minimal reward gap \(\delta\).
    Set budget \(T=n\).
    Initialize new epoch: Query \(K\) new arms; call it consideration set \(\mathcal{A}=\{1, \ldots, K\}\).
    Set exploration duration \(L=\left\lceil 2 \underline{\delta}^{-2} \log n\right\rceil ; m \leftarrow \min (L,\lfloor T / K\rfloor)\).
    Play each arm in \(\mathcal{A} m\) times; observe rewards \(\left\{\left(X_{1, j}, \ldots, X_{K, j}\right): j=1, \ldots, m\right\}\).
    Update budget: \(T \leftarrow T-K m\).
    if \(\exists a, b \in \mathcal{A}, a<b\) s.t. \(\left|\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)\right|<\underline{\delta} m\) then
    Permanently discard \(\mathcal{A}\), and repeat from step (3).
    else
            Permanently commit to \(\operatorname{arm} a^{*} \in \arg \max _{a \in \mathcal{A}}\left\{\sum_{j=1}^{m} X_{a, j}\right\}\).
```

Setup 1 [Figures 2.1, 2.2 and 2.3]. In this setting, we consider $K=2$ with $\alpha_{1}=0.5$, i.e., two equiprobable arm-types, characterized by Bernoulli(0.6) and Bernoulli(0.4) rewards. Via this setup, we intend to illustrate the difference between the empirical performance achievable in the countable-armed setting vis-à-vis its traditional two-armed counterpart. Refer to Figure 2.1. The red curve indicates the empirical performance of ALG3 in this setting. For reference, the blue one shows the empirical performance of UCB1 [10] in a two-armed bandit with Bernoulli(0.6) and Bernoulli(0.4) rewards; the green curve indicates the best achievable instance-dependent regret [3] in said two-armed configuration. As expected, the regret of ALG3 is inflated relative to UCB1. This is owing to the $\beta_{\delta, 2} \leqslant 1$ factor present in the denominator of ALG3's upper bound; characterization of the sharpest lower bound on the probability in (2.2) (see Proposition 1) is challenging owing to the limited theoretical tools available to this end and we leave it as an open problem at the moment. Figure 2.2 shows the empirical performance of the algorithms proposed in this paper as well as Sampling-UCB initialized with $\alpha_{1}=1 / 2$ and ETC- $\infty$ (2) initialized with $\underline{\delta}=\delta / 2=$ 0.1. Evidently, the (adaptive) explore-then-commit approach in $\operatorname{ALG} 2(n)$ outperforms the prespecified exploration schedule-based approach of $\operatorname{ALG1}(n)$, and performs almost as good as the gap-aware approach in ETC- $\infty$ (2). While Sampling-UCB outmatches all explore-then-commit styled approaches, the best performing algorithm is ALG3. Surprisingly, this is despite the fact that the theoretical performance bounds for $\operatorname{ALG} 2(n)$ and $\operatorname{ALG} 3$ are identical (modulo numerical multiplicative constants) when $K=2$ and $\alpha_{1} \leqslant 0.5$ (see Theorem 10 and 11). A similar hierarchy in performances is also observable in Figure 2.3, which corresponds to a slightly "easier" instance with $\delta=0.4$ (as opposed to 0.2 ) and equiprobable Bernoulli(0.9) and Bernoulli(0.5) rewards.


Figure 2.1: $K=2$ and $\boldsymbol{\alpha}=(1 / 2,1 / 2)$ : Achievable regret in 2-CAB vis-à-vis 2-MAB.


Figure 2.2: $K=2$ and $\alpha=(1 / 2,1 / 2)$ : An instance with Bernoulli 0.6, 0.4 rewards.

Setup 2 [Figure 2.4]. Here, we consider a setting with $K=3$ arm-types characterized by Bernoulli rewards with means $0.9,0.5,0.1$, each occurring with probability $1 / 3$. We compare the performance of $\operatorname{ALG} 1(n), \operatorname{ALG} 2(n)$ and $\operatorname{ALG} 4$ with $\operatorname{ETC}-\infty(3)$ initialized with $\underline{\delta}=\delta / 2=0.2$, and Sampling-UCB initialized with $\alpha_{1}=1 / 3$. It is noteworthy that despite ALG4's significantly superior empirical performance relative to aforementioned algorithms, only a weak $o(n)$ theoretical guarantee on its regret is currently available (see Theorem 12) due to reasons discussed earlier in the paper. Investigating best achievable rates under ALG4 is an area of active research at the moment.


Figure 2.3: $K=2$ and $\alpha=(1 / 2,1 / 2)$ : An instance with Bernoulli $0.9,0.5$ rewards.


Figure 2.4: $K=3$ and $\boldsymbol{\alpha}=(1 / 3,1 / 3,1 / 3)$ : Instance with Bernoulli $0.9,0.5,0.1$ rewards.

### 2.6 Concluding remarks, extensions, and open problems

This chapter summarizes the contributions in $[21,48]$ and provides a first-order characterization of the complexity of the $K$-typed countable-armed bandit problem with matching lower and upper bounds for $K=2$. For $K>2$, we establish an instance-dependent upper bound of $O\left(K^{3} \bar{\Delta} / \delta^{2} \log n\right)$ and show that the scaling w.r.t. $K$ cannot beat $\Omega(K \log K)$; the latter property differentiates this setting fundamentally from the classical $K$-armed problem. Another key takeaway from our work is that achievable regret in this setting only has a second-order dependence on the reservoir distribution, i.e., dependence on $\alpha$ only manifests through sub-logarithmic terms (see Theorem 10 and 11). Although this work is predicated on countably many arms, our algorithms can easily be adapted to settings with a large but finite number of arms. For example, the result on second-order dependence w.r.t. $\alpha$ has profound implications for the $N$-armed bandit problem with $K$ arm-types, where each type is characterized by a unique mean reward. A naive implementation of standard MAB algorithms in this setting will result in a regret that scales linearly with $N$. Instead, one can simulate a $K$-typed reservoir over the $N$ arms and deploy ALG2(n) to achieve an $O\left(K^{3}\right)$ scaling of the leading term; if $K \ll N$, performance improvement can be substantial vis-à-vis naive MAB algorithms. Another important direction concerns adaptivity to $K$ : This paper provides algorithms that adapt to $\alpha$ assuming perfect knowledge of $K$; performance characterization given only an approximation thereof remains an open problem.

An extension of the countable-armed framework introduced in this chapter has been used as a basal motif in sequential assignment and dynamic matching problems arising in settings such as online labor markets (see Chapter 3). Another extension, pursued in Chapter 4, studies a setting where the arm-reservoir distribution is endogenous.

# Chapter 3: Countable-armed bandits: An application to matching markets 

### 3.1 Introduction

The problem of sequentially matching "jobs" to "workers" under uncertainty forms the bedrock of many modern operational settings, especially in the online gig economy, see, e.g., applications such as Amazon Mechanical Turk, TaskRabbit, Jobble, and the likes. A simpler instance of the problem dates back to [49] where it is referred to as the sequential stochastic assignment problem (SSAP). A fundamental issue in such settings is that the platform typically is oblivious (at least initially) to the skill proficiencies of individual workers for specific job categories. This complexity is further compounded by the large number of workers usually present on such platforms, tantamount to prohibitively large experimentation costs associated with acquisition of granular information at the level of an individual worker. This issue is commonly mitigated by exploiting structure in the problem (if any), or by positing distributional assumptions on the population of available workers, e.g., workers may be drawn from some distribution $\mathcal{D}$ satisfying certain context-specific desiderata. Such distributional assumptions are vital to designing efficient algorithms for these systems, and as such, traditional literature has largely relied on the availability of ex ante knowledge of $\mathcal{D}$ or certain key aspects thereof (refer to the literature review below).

Key research question. An important characteristic of the gig economy is that the population of workers may undergo distributional shifts over the course of the platform's planning horizon. These effects may, many a time, fail to register in a timely manner; as a result, there may be delays in tailoring appropriately the matching algorithm (calibrated typically using available distributionlevel information) to the changed environment. This has the potential to cause revenue losses as well as catalyze endogenous worker attrition. Such exigencies necessitate designing algorithms that are agnostic to $\mathcal{D}$ and whose performance is robust to plausible realizations thereof.

The model at a glance. We consider a finite set of possible job-types (denoted by $\mathcal{J}$ ), an assumption we deem appropriate for settings such as those discussed above. In addition, we model workers as exhibiting discrete skill-levels (aka worker-types), indexed by $\left\{1, \ldots, K_{j}\right\}$, w.r.t. each job-type $j \in \mathcal{J}$, and assume that $\left(K_{j}: j \in \mathcal{J}\right)$ is known a priori. It is not unreasonable to make this assumption since it is common, in practice, for platforms to deploy pilot experiments prior to the actual matching phase in order to gather sufficient information on key primitives such as the size and stability of low-dimensional sub-population clusters, if any exist; one can therefore safely assume in settings where such structure exists that $\left(K_{j}: j \in \mathcal{J}\right)$ is well-estimated a priori.

While the demand is constituted by sequential job-arrivals (possibly in batches of stochastic size and composition), we posit availability of an unlimited number of workers on the supply side. This feature encapsulates the choice overload phenomenon characteristic of many large market settings where workers are available in a large number relative to the platform's planning horizon. To our best knowledge, extant literature on matching under uncertainty is largely limited to "finite" markets (see the literature review below), and therefore fails to accommodate this important practical consideration. In our setting, the population of workers, albeit large, is governed by a finitely supported distribution that controls the proportion of each worker-type. Specifically, the $K_{j}$ distinct worker-types w.r.t. job-type $j$ are distributed according to $\alpha_{j}:=\left(\alpha_{i, j}: i=1, \ldots, K_{j}\right)$, where $\sum_{i=1}^{K_{j}} \alpha_{i, j}=1$. We note that this is one possible model of a matching market that is closer in spirit to SSAP [49]; it differs from other models in the matching literature (refer to the literature review below) in that it tries to capture a salient aspect of large markets, viz., choice overload, as opposed to aspects such as competition and congestion best elucidated via traditional "finite" market models.

The platform's goal is to maximize its expected cumulative payoffs over a sequence of $n$ rounds of matching, subject to worker-types w.r.t. job-types and their distributions $\left\{\alpha_{j}: j \in \mathcal{J}\right\}$, as well as mean payoffs for possible worker-job type-pairs being latent attributes. As is the norm in settings with incomplete information and imperfect learning, we reformulate this objective as minimizing the expected cumulative regret relative to an oracle that is privy to aforementioned primitives.

On the complexity of the problem. With a unique job-type, say $\mathcal{J}=\left\{j_{0}\right\}$, and only one job arriving per period, note that the ensuing allocation problem reduces to the countable-armed bandit problem introduced in Chapter 2. The task then reduces to cleverly aggregating such countablearmed bandits to solve the original matching problem.

Literature review. Our problem is situated close to a recent line of work on dynamic matching under uncertainty. This stream of literature, by and large, considers an archetypal decentralized matching problem under uncertainty in preferences where a heterogeneous collection of jobs (represented by nodes on one side of a bipartite graph) must be matched to workers (the other side of the graph) with unknown or noisy preferences over jobs (see, e.g., [50] and references therein). The matching proceeds iteratively in rounds in a way that meets certain stability criteria at all times as well as ensures that the true preferences are "learnt" at a regret-optimal rate. Such models, however, are fundamentally distinct from ours in that their learning problems are posited over a finite set of workers, which allows for sufficient exploration of each; this would be infeasible in our setting owing to a "large" population thereof.

Another related setting is studied in [37] where a centralized steady-state model with endogenous workers is considered. The key technical innovation in this work lies in the way polytope capacity constraints are handled via shadow prices to create essentially an unconstrained learning problem that may be solved rate-optimally using conventional heuristics. However, this necessitates ex ante knowledge of $\left\{\boldsymbol{\mu}_{\boldsymbol{j}}, \boldsymbol{\alpha}_{\boldsymbol{j}}: j \in \mathcal{J}\right\}$ among other problem primitives. Our model, on the other hand, has a richer learning component that is challenging to address as it is, absence of capacity constraints notwithstanding. Our contribution lies in establishing fundamental achievability results for the learning problem via reduction to a countable-armed bandit setting, and in the design of novel rate-optimal algorithms that adapt to the problem primitives online. We leave it to future work to incorporate into our model more realistic features such as capacity constraints, strategic interactions between the platform and workers, etc.

### 3.2 Problem formulation

Job-arrival process. The platform faces an arrival stream of jobs (i.i.d. in time) given by $\left\{\left(\Lambda_{j, t}: j \in \mathcal{J}\right): t=1,2, \ldots\right\}$, where $\mathcal{J}$ is finite and $\Lambda_{j, t}$ is the number of type $j$ jobs arriving at time $t$. Types and multiplicities of jobs are perfectly observable upon arrival. We assume that there exists some finite constant $M>0$ satisfying $\mathbb{P}\left(\max _{j \in \mathcal{J}} \sup _{t \geqslant 1} \Lambda_{j, t} \leqslant M\right)=1$. Note that our algorithms do not require knowledge of $M$; the assumption only serves to simplify analysis and can be relaxed.

Supply of workers. We assume that workers are distributed on the unit interval $[0,1]$ according to some probability distribution $\mathcal{D}$ that is absolutely continuous w.r.t. the Lebesgue measure on $[0,1]$. Associated with each job-type $j \in \mathcal{J}$, there exists a permutation $\sigma_{j}:=$ $\left\{\sigma_{j}(i): i=1, \ldots, K_{j}\right\}$ of $\left\{1, \ldots, K_{j}\right\}$, and a sequence of thresholds $0=: \lambda_{0, j}<\lambda_{1, j}<\ldots<$ $\lambda_{K_{j}-1, j}<\lambda_{K_{j}, j}:=1$ partitioning the unit interval into $K_{j}$ disjoint sub-intervals. We posit a payoff model whereby a worker $x \in\left(\lambda_{i-1, j}, \lambda_{i, j}\right)$ (for some $i \in\left\{1, \ldots, K_{j}\right\}$ ) generates a stochastic reward with mean $\mu_{\sigma_{j}(i), j}$ upon match with a type $j$ job; it is assumed that the $K_{j}$ mean rewards adhere to the strict order $\mu_{1, j}>\ldots>\mu_{K_{j}, j}$. We define $\alpha_{i, j}:=\mathbb{P}\left(X \in\left(\lambda_{\iota(i, j)-1, j}, \lambda_{\iota(i, j), j}\right)\right)$, where $X \sim \mathcal{D}$ and $\iota(i, j) \in\left\{1, \ldots, K_{j}\right\}$ is the unique element satisfying $\sigma_{j}(\iota(i, j))=i$, as the probability that a worker sampled at random from $\mathcal{D}$ (equivalently, from the population), is $i^{\text {th }}$ best for job-type $j$ (generates mean reward $\mu_{i, j}$ ); such a worker is said to have type $i$ w.r.t. job-type $j$. Thus, a type 1 worker w.r.t. job-type $j$ is optimal for jobs of type $j$. Note that the model allows for staggered optimality of worker-types; see Figure 3.1.


Figure 3.1: Possible distribution of worker-types for $\mathcal{J}=\{\mathbf{1 , 2}\}$ and $\boldsymbol{K}_{\mathbf{1}}=\boldsymbol{K}_{\mathbf{2}}=\mathbf{2}$. The darker shades represent type 1 (optimal) workers while the lighter shades represent type 2 (inferior) workers w.r.t. each job-type in $\mathcal{J}$. In this example, no worker can simultaneously be optimal for both job-types.

High-level description of the matching problem. Each arriving job may be matched one-to-one to a worker from the available supply. Each match takes one period for execution, it is
therefore possible to match jobs arriving in consecutive periods to the same worker. Matched jobs leave the system upon completion and the platform receives a stochastic reward for each completed job; a job that remains unmatched drops out instantaneously. The platform has information neither on individual worker-types w.r.t. job-types nor on their supply distribution, however, it has perfect knowledge of $\left(K_{j}: j \in \mathcal{J}\right)$. Subject to this premise, the platform must match incoming jobs to workers in a way that maximizes its expected cumulative payoffs over $n$ rounds of matching.

Adaptive control. For any job that arrives at time $t$, the platform can match it to: (i) a worker that has matched before, (ii) a new worker (one without any history of matches) sampled from the population, or (iii) no worker (job is dropped). To this end, a policy $\pi:=\left(\pi_{1}(\cdot, \cdot), \pi_{2}(\cdot, \cdot), \ldots\right)$ is an adaptive rule that prescribes the allocation $\pi_{t}(\cdot, \cdot)$ at time $t$. Specifically, $\pi_{t}(j, k)$ encodes the worker that should match with the $k^{\text {th }}$ job of type $j$ arriving at time $t$ (provided there are at least $k$ job-arrivals of type $j$ at $t$ and the $k^{\text {th }}$ job is not dropped). Upon match, a [0, 1]-valued stochastic reward with mean $\mu_{\kappa_{j}\left(\pi_{t}(j, k)\right), j}$ is realized, where $\kappa_{j}\left(\pi_{t}(j, k)\right) \in\left\{1, \ldots, K_{j}\right\}$ denotes the type of worker $\pi_{t}(j, k)$ w.r.t. job-type $j$. The realized rewards are independent across matches and in time.

Platform's objective. The goal of maximizing the expected cumulative payoffs over $n$ rounds is converted to minimizing the expected regret relative to a clairvoyant policy that prescribes an "optimal" match for each arriving job. We are thus interested in the following optimization problem

$$
\begin{equation*}
\inf _{\pi \in \Pi} \mathbb{E} R_{n}^{\pi}:=\inf _{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{n} \sum_{j \in \mathcal{J}: \Lambda_{j, t}} \sum_{k=1}^{\Lambda_{j, t}}\left(\mu_{1, j}-\mu_{\kappa_{j}\left(\pi_{t}(j, k)\right), j}\right)\right] . \tag{3.1}
\end{equation*}
$$

Here, $\Pi$ is the class of non-anticipating policies, i.e., $\pi_{t+1}(\cdot, \cdot)$ is adapted to $\mathcal{F}_{t}$ for all $t \in\{0,1, \ldots\}$, where $\mathcal{F}_{t}:=\sigma\left\{\left(\boldsymbol{\Lambda}_{s}, \boldsymbol{\pi}_{s}, \boldsymbol{r}_{s}\right): s=1, \ldots, t\right\}$ denotes the natural filtration at time $t$. Here, $\boldsymbol{\Lambda}_{s}:=$ $\left(\Lambda_{j, s}: j \in \mathcal{J}\right), \boldsymbol{\pi}_{s}$ is the set of matches implemented at time $s$ and $\boldsymbol{r}_{s}$ is the set of collected rewards. The expectation in (3.1) is w.r.t. the randomness in job-arrivals, worker supply, policy, and rewards.

Going forward, we will adopt standard terminology from the multi-armed bandit literature and refer to workers as "arms" and jobs as "pulls" interchangeably.

### 3.3 Designing adaptive policies for matching

The approach we adopt in this paper directly addresses the fact that there is an unlimited supply of available workers at all times. A natural design then is to tailor sub-routines specific to jobtypes in $\mathcal{J}$ and instantiate them at the first arrival of each type. Specifically, if jobs of type $j$ arrive at $\left\{t_{1}, t_{2}, \ldots\right\}$, then the platform should call the sub-routine specific to job-type $j$ only at aforementioned times, independent of other job-arrivals. This leads to the meta-algorithm MATCH (see Algorithm 9) for the matching problem. In what follows, ALG refers to an arm-allocation rule w.r.t. a fixed job-type that prescribes one arm upon each invocation. ALG can be thought of as a horizon-free sampling strategy for a countably many-armed bandit problem with one pull per period. When multiple jobs (say $L$ ) of the same type (say $j$ ) arrive at the same time, we instantiate (if necessary) new parallel threads of ALG specific to job-type $j$ to ensure that the demand is fully met. In the following, $L_{j}$ denotes the running count of parallel threads of ALG for type $j$ jobs.

```
Algorithm 9 MATCH: A meta-algorithm for the matching problem
    Input: (i) \(\mathcal{J}\), (ii) \(\left(K_{j}: j \in \mathcal{J}\right)\), and (iii) ALG.
    Initialization: Set \(L_{j}=0\) for each \(j \in \mathcal{J}\).
    for \(t \in\{1,2, \ldots\}\) do
        for \(j \in \mathcal{J}\) do
            if \(\Lambda_{j, t} \geqslant 1\) then
            if \(\Lambda_{j, t} \leqslant L_{j}\) then
                            Match the \(\Lambda_{j, t}\) type \(j\) jobs to the first \(\Lambda_{j, t}\) threads of ALG for type \(j\) jobs.
            else
                    Match the first \(L_{j}\) type \(j\) jobs to the \(L_{j}\) available threads of ALG for type \(j\) jobs.
                    Instantiate \(\Lambda_{j, t}-L_{j}\) new threads of ALG for the remaining \(\Lambda_{j, t}-L_{j}\) jobs.
                    Update \(L_{j} \leftarrow \Lambda_{j, t}\).
```

Discussion of MATCH. An immediate observation from Algorithm 9 is that ALG ought to be anytime, i.e., it should not depend on the horizon of play since the number of jobs of each type
arriving over the platform's planning horizon is not known a priori. Keeping this objective in mind, we shift our focus to designing an arm-allocation rule ALG w.r.t. a fixed job-type, say type $j$, that: (i) prescribes one pull per period, (ii) depends only on $K_{j}$, (iii) is adaptive to the mean reward vector $\boldsymbol{\mu}_{\boldsymbol{j}}$ and the supply distribution $\boldsymbol{\alpha}_{\boldsymbol{j}}$, and (iv) is horizon-free. Upon successful design of ALG, its composition with MATCH will transfer learning guarantees to the original matching problem.
3.3.1 Shifting focus to adaptive sequential sampling strategies tailored to a specific job-type

Going forward, we will assume that jobs belong to a common fixed type and arrive one at a time. With slight abuse of notation, the supply of available workers is characterized by $K$ workertypes with distinct means $\boldsymbol{\mu}:=\left(\mu_{i}: i=1, \ldots, K\right)$ adhering to $\mu_{1}>\ldots>\mu_{K}$. The maximal and minimal sub-optimality gaps are given by $\bar{\Delta}:=\mu_{1}-\mu_{K}$ and $\underline{\Delta}:=\mu_{1}-\mu_{2}$ respectively, and the minimal reward gap is $\delta:=\min _{1 \leqslant i<i^{\prime} \leqslant K}\left(\mu_{i}-\mu_{i^{\prime}}\right)$. The distribution of worker-types is denoted by $\boldsymbol{\alpha}:=\left(\alpha_{i}: i=1, \ldots, K\right)$, where $\alpha_{i}$ is the probability of sampling a type $i$ arm (with mean $\left.\mu_{i}\right)$ from the population. The decision maker knows $K$ but is oblivious to ( $\boldsymbol{\mu}, \boldsymbol{\alpha}, n$ ), $n$ being the learning horizon.

Note that the aforementioned setting is that of the countable-armed bandit problem introduced in Chapter 2 in which ALG2 (see Algorithm 5) achieves rate-optimal regret with a second-order dependence on the reservoir distribution (see Theorem 10).

### 3.3.2 Transferring learning guarantees to the matching problem

A horizon-free version of ALG2 that preserves its logarithmic regret guarantee may be obtained by passing it to the $\mathcal{D T}$ operator [51] with a doubling sequence of $T_{i}=2^{2^{i}}$.

Theorem 13 (Achievable performance under MATCH $\circ \mathcal{D T}$ (ALG2)) Denote by $\pi$ the composition of MATCH with $A L G=\mathcal{D T}$ (ALG2). Then, after any number $n \geqslant 1$ of rounds, one has

$$
\begin{equation*}
\mathbb{E} R_{n}^{\pi} \leqslant C M \sum_{j \in \mathcal{J}}\left[\frac{K_{j}^{3} \bar{\Delta}_{j}}{\beta_{\delta_{j}, K_{j}}}\left(\frac{\log n}{\delta_{j}^{2}}+\frac{\log \log (n+2)}{K_{j}!\prod_{i=1}^{K_{j}} \alpha_{i, j}}\right)\right], \tag{3.2}
\end{equation*}
$$

where $\bar{\Delta}_{j}:=\mu_{1, j}-\mu_{K_{j}, j}, \delta_{j}:=\min _{1 \leqslant i<i^{\prime} \leqslant K_{j}}\left(\mu_{i, j}-\mu_{i^{\prime}, j}\right), \beta_{\delta_{j}, K_{j}}$ is as defined in (2.2) with $\delta \leftarrow \delta_{j}$, $K \leftarrow K_{j}$, and $C$ is some absolute constant.

Discussion. The foremost noticeable aspect of Theorem 13 is that achievable regret depends on $\left\{\alpha_{j}: j \in \mathcal{J}\right\}$ (collection of worker-type distributions w.r.t. job-types), surprisingly, only through $o(\log n)$ terms. Among other things, characterizing the minimax complexity of this problem setting remains a challenging open problem in light of the unconventional multiplicative factors in (3.2).

Proof of Theorem 13. Using Theorem 7 of [51] for a doubling sequence of $T_{i}=2^{2^{i}}$, together with Theorem 10 (see Chapter 2), one obtains that $\mathbb{E} R_{n}^{\mathcal{D T}(A L G 2)} \leqslant \frac{C K^{3} \bar{\Delta}}{\beta_{\delta, K}}\left(\frac{\log n}{\delta^{2}}+\frac{\log \log (n+2)}{K!\prod_{i=1}^{K} \alpha_{i}}\right)$ for some absolute constant $C$. In the matching problem, note that there are exist at most $M|\mathcal{J}|$ active threads of $\mathcal{D T}$ (ALG2) at any time. Since $\mathcal{D T}$ (ALG2) is horizon-free, the regret incurred is dominated by that in the scenario where $M|\mathcal{J}|$ threads are active at each $t \in\{1, \ldots, n\}$. The assertion is now immediate.

### 3.4 Concluding remarks

This chapter summarizes the contributions in [52] and provides a stylized model for data-driven matching in settings such as online labor markets. There remain a lot of practical considerations, e.g., queuing, incentives, competition, and congestion that our model currently does not address; each of these aspects independently constitutes an important direction for future work.

# Chapter 4: Countable-armed bandits with dynamic arm-acquisition costs: Towards a non-stationary arm-reservoir 

### 4.1 Introduction

Cost of reservoir access. In multi-armed bandit (MAB) settings where large action space is a defining characteristic, the decision maker may experience elevated costs of acquiring new arms as time progresses. This can be viewed in light of plausible deterioration in the "quality" of the reservoir either with increasing number of queries or with time. The former can be interpreted as a Lagrangian relaxation to an optimization problem with capacity constraints on reservoir queries; a setting potentially of interest to online resource allocation problems under costly resource acquisition. The latter aspect has connotations related to market churn. For example, the relative abundance of agents of different types in a market may undergo temporal variations over the platform's planning horizon. This is the case, for example, in online service-platforms, where arms may represent agents capable of abandoning the platform if kept idle for protracted durations. These effects may catalyze an agent-departure process and a temporal non-stationarity that may seriously hinder the decision maker's ability to discern "good" arms from "inferior" ones. There has been significant recent interest in non-stationary bandit models, however, most of the literature is limited largely to non-stationarities in rewards, and antecedents on non-stationarities in arm-reservoirs are markedly absent (see the literature review in §4.1.1). In this paper, we provide the first systematic treatment of this aspect by investigating the statistical limits of learning under endogenous variations in the arm-reservoir distribution.

Stylized model. To distill key insights, it will be convenient to focus on the paradigmatic case of exactly two arm-types in the reservoir; this serves to highlight statistical idiosyncrasies of the problem without unnecessary mathematical detail. The complexity of the static version
of the problem is driven by three primitives: (i) the gap $\Delta$ between the mean rewards associated with the two arm-types; (ii) the fraction $\alpha$ of the "optimal" arm-type in the reservoir; and (iii) the horizon of play $n$. The feature of costly arm-acquisitions is incorporated by endogenizing $\alpha$; consequently, its evolution is given by the stochastic process $(\alpha(t): t=1,2, \ldots)$. Thus, at any time $t$, the probability that a query of the arm-reservoir returns an arm of the optimal type is given by (the sample path-dependent random variable) $\alpha(t)$. As we shall later see, the interesting regime is where $\alpha(t)$ becomes vanishingly small as $t$ grows; this captures departure of optimal (unexplored) arms from the reservoir, and abstracts out key characteristics of many online platforms that serve a large population of impatient agents.

Challenges due to costly reservoir access. It is non-trivial to design "good" policies that are reservoir distribution-agnostic as well as gap-adaptive. To see this, consider the simplest scenario where $\alpha(t)=c$ (constant), and the decision maker is endowed with ex ante knowledge of $c$ as well as the horizon $n$. A natural heuristic in this setting is to query $\Omega\left(c^{-1} \log n\right)$ arms upfront (this would guarantee with probability $\Omega(1 / n)$ the existence of at least one optimal arm among the queried ones), and subsequently deploy a conventional bandit algorithm such as Thompson Sampling [13] or UCB [10] on the collected set of arms. One can quite easily show that such an approach will incur poly-logarithmic regret in $n$. On the other hand, it is possible to achieve $\tilde{O}\left((c \Delta)^{-1} \log n\right)$ regret (up to poly-logarithmic factors in $\Delta^{-1}$ ) using a more sophisticated policy [43]. Regardless, none of these approaches utilize the fact that there exist exactly $K$ arm-types in the reservoir. By leveraging this knowledge, $c$-dependence of the aforementioned upper bound can, in fact, be relegated to sub-logarithmic terms, surprisingly, in a manner that is adaptive to $c$ [21]. It is noteworthy, however, that the performance of such approaches is fragile w.r.t. the premise that $\alpha(t)$ remains bounded away from 0 by some problem-independent constant $c$ (which may or may not be known a priori) at all times. Naturally, endogenous variations potentially causing $\alpha(t)$ to vanish in $t$ will only exacerbate the problem, and it remains unclear if it is even possible in this setting to achieve sub-linear regret relative to the classical full-information benchmark that prescribes pulling an optimal arm at each $t$.

Contributions. We first derive a necessary condition for "complete learning" when the evolution of $\alpha(t)$ is independent of the decision maker's actions; specifically, $\sum_{t=1}^{\infty} \alpha(t)=\infty$ is necessary for achieving sub-linear regret relative to an oracle that knows the identities of optimal arms ex ante (Theorem 14). We also establish its near-tightness in that a slightly stronger version of said condition is sufficient for a gap-aware policy based on the Explore-then-Commit principle to achieve poly-logarithmic instance-dependent regret in the problem (Theorem 15). In addition, we discuss a novel gap-adaptive policy based on the UCB principle that achieves a polynomial regret in the same regime (Theorem 16). We then consider the setting where $\alpha(t)$ is endogenous (policy-dependent), and characterize matching necessary and sufficient conditions (up to leading order terms) for asymptotic-optimality of aforementioned policies.

Before proceeding with a formal description of our model and discussion of results, we provide a brief overview of related literature below.

### 4.1.1 Related literature

Bandits with state-dependent rewards. The earliest work on MAB problems involving endogenous arm-reward distributions dates back to the seminal work [53] which studied finite-state Markovian bandits. In this model, the state of an arm only changes upon execution of a pull while remaining unchanged otherwise, thus prompting the name resting bandits; the celebrated Gittins index policy is well-known to maximize the infinite-horizon discounted cumulative expected reward in this setting. In contrast, in the so-called restless bandits formulation [54], the states of all the arms may change simultaneously irrespective of which arms are pulled; in addition, this formulation permits pulling any fixed number of arms in each period. Subsequent works such as [55] focus on heuristics that are optimal in an asymptotic regime where the number of arms pulled in each period scales linearly with the total number of arms. More recently, a finite-horizon variant of the restless bandits problem was studied in [56] under a similar scaling; see [57, 58, 59], etc., for a survey of other well-studied variations.

Our work is quite distinct from this strand of literature: (i) asymptotic analyses in aforemen-
tioned papers are w.r.t. the number of arms, not the horizon of play; (ii) the number of arms pulled in each period is fixed at 1 in our problem setting and does not scale with the total number of arms; (iii) most importantly, cited references formulate the problem as a Markov decision process assuming full knowledge of the transition kernels. In contrast, we consider a learning theoretic formulation where the decision maker is oblivious to the statistical properties of reward distributions as well as the nature of endogeneity in the arm-reservoir.

Bandits with non-stationary rewards. This line of work focuses on policies that minimize the expected cumulative regret relative to a dynamic oracle that plays at each time $t$ an arm with the highest mean reward at $t$. Some of the early work in this area is premised on a formulation in which the identity of the best arm may change a finite number of times adversarially during the horizon of play, see, e.g., [60]. While other works such as [61] study specific models of temporal variation where, for example, rewards evolve according to a Brownian motion, much of the traditional literature is limited, by and large, to a finite number of changes in the mean rewards; see [62] and references therein. Subsequently, a unified framework for studying aforementioned problem classes was provided in [63] by introducing a variation budget to bound the evolution of mean rewards over the horizon of play. Several other forms of non-stationarity have also been studied in the literature; these include formulations with rotting [64], recharging [65], and delaydependent rewards [66], among others (see [67] for a survey).

Aforementioned works are largely limited to the study of finite-armed bandit problems where non-stationarity can be ascribed to changes in arm-means. In contrast, our work differs fundamentally in that it is premised on an infinite-armed formulation with non-stationarity attributable to an endogenous arm-reservoir. In a nutshell, while preceding work focuses on distributional shifts in rewards given a fixed set of arms, we propose a new paradigm where the arm-reservoir itself undergoes distributional shifts owing to possible "leakages," which is functionally a very different concept.

Bandits with infinitely many arms. These problems involve settings where an unlimited supply of arms is governed by some fixed distribution over an uncountable set of arm-types (possible
mean rewards); a common reward statistic (usually the mean) uniquely characterizes each armtype. The infinite-armed bandit problem traces its roots to [38] where it was studied under the Bernoulli reward setting with means distributed Uniformly on [ 0,1 ]. More general reward and reservoir distributions on [0, 1] have also been studied in subsequent works, see, e.g., [39, 40, 41, 42]. Our model differs from this line of work in that we only assume the reservoir to be finitely supported and posit no distributional knowledge thereof, unlike cited references. Furthermore, the distribution of arm-types is allowed to vary endogenously over the problem horizon in a manner that may be unknown ex ante.

### 4.1.2 Outline of this chapter

A formal description of the model is provided in $\S 4.2$. $\S 4.3$ discusses reservoir distributionagnostic algorithms; two natural modes of endogeneity in the reservoir distribution are discussed in §4.4 along with corresponding guarantees on achievable performance. $\S 4.5$ provides concluding remarks. Proofs and ancillary results are delegated to Appendix C.

### 4.2 Problem formulation

Primitives. There are finitely many possible arm-types in the reservoir denoted by the collection $\mathcal{K}$; the vector $\boldsymbol{\mu}:=\left(\mu_{k}: k \in \mathcal{K}\right)$ characterizes the mean reward (pairwise distinct) associated with each. The decision maker (DM) only knows the cardinality of $\mathcal{K}$; for simplicity of exposition, we assume $|\mathcal{K}|=2$ in this work and index the two arm-types by " 1 " and "2," i.e., $\mathcal{K}=\{1,2\}$. Without loss of generality, arm-type 1 is assumed "optimal" with a gap (or separation) of $\Delta:=\mu_{1}-\mu_{2}>0$ from the "inferior" type (arm-type 2); as we shall later see, $\Delta$ is an important driver of the problem's statistical complexity. DM must play one arm at each time $t \in\{1, \ldots, n\}$, where $n$ denotes the horizon of play.

Reservoir access. The collection of arms to have been played at least once until time $t$ (inclusive) is denoted by $\mathcal{I}_{t}$ (with $\mathcal{I}_{0}:=\phi$ ). The set of available actions at $t$ is given by $\mathcal{A}_{t}=\mathcal{I}_{t-1} \cup\left\{\right.$ new $\left._{t}\right\} ;$ DM must either play an arm from $\mathcal{I}_{t-1}$ or select the action "new ${ }_{t}$ " which corresponds to play-
ing a new arm queried from the reservoir. A newly queried arm at time $t$ is optimal-typed with probability $\alpha(t)$ and inferior-typed otherwise ( $\alpha(t)$ may potentially be random and endogenous, unbeknown to DM; a detailed discussion is deferred to §4.4). Arm-types are private attributes and remain unobservable throughout. Upon pulling an arm labeled $i$ (henceforth referred to as arm $i$ ) for the $j^{\text {th }}$ time, DM observes a $[0,1]$-valued stochastic reward denoted by $X_{i, j}$. The realized rewards are independent across arms and time, and mean-preserving in time (not necessarily identically distributed) keeping the arm fixed.

Admissible controls. A policy $\pi:=\left(\pi_{1}, \pi_{2}, \ldots\right)$ is an adaptive allocation rule that prescribes at time $t$ an action $\pi_{t}$ (possibly randomized) from $\mathcal{A}_{t}$. The collection of all observable information until $t$ is given by the natural filtration $\mathcal{F}_{t}:=\sigma\left\{\left(\pi_{s}\right)_{1 \leqslant s \leqslant t},\left\{\left(X_{i, j}\right)_{1 \leqslant j \leqslant N_{i}(t)}: i \in \mathcal{I}_{t}\right\}\right\}$ (with $\mathcal{F}_{0}:=$ $\phi)$, where $N_{i}(t)$ indicates the number of times arm $i$ is pulled until $t$. The cumulative regret of $\pi$ after $n$ plays is given by $R_{n}^{\pi}:=\sum_{t=1}^{n}\left(\mu_{1}-X_{\pi_{t}, N_{\pi_{t}}(t)}\right)$ and its cumulative pseudo-regret by $\mathcal{R}_{n}^{\pi}:=$ $\sum_{t=1}^{n}\left(\mu_{1}-\mu_{\kappa\left(\pi_{t}\right)}\right)$, where $\kappa\left(\pi_{t}\right) \in \mathcal{K}$ encodes the type of arm $\pi_{t}$; note that both regret as well as pseudo-regret are sample path-dependent by definition. DM is interested in the following stochastic minimization problem

$$
\begin{equation*}
\inf _{\pi \in \Pi} \mathbb{E} R_{n}^{\pi}=\inf _{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{n}\left(\mu_{1}-X_{\pi_{t}, N_{\pi_{t}}(t)}\right)\right]=\inf _{(\dagger)} \mathbb{E}\left[\sum_{t=1}^{n}\left(\mu_{1}-\mu_{\kappa\left(\pi_{t}\right)}\right)\right]=\inf _{\pi \in \Pi} \mathbb{E} \mathcal{R}_{n}^{\pi}, \tag{4.1}
\end{equation*}
$$

where the infimum is over the class $\Pi$ of non-anticipating policies, i.e., $\pi_{t}: \mathcal{F}_{t-1} \rightarrow \mathcal{P}\left(\mathcal{A}_{t}\right) ; t \in$ $\{1,2, \ldots\}\left(\mathcal{P}\left(\mathcal{A}_{t}\right)\right.$ denotes the probability simplex on $\left.\mathcal{A}_{t}\right)$, the expectations are w.r.t. all the possible sources of stochasticity in the problem (rewards, policy, and the reservoir distribution), and ( $\dagger$ ) holds since cumulative regret and pseudo-regret are equal in expectation in the stochastic bandits setting (follows from the Tower property of expectations).

### 4.3 Distribution-agnostic policies

As discussed in the introduction, our goal in this work is to investigate achievable regret under algorithms that are agnostic to the reservoir distribution. To this end, we discuss two
such algorithms; one based on the forced-exploration principle, and the other on optimism-underuncertainty.

### 4.3.1 A fixed-design policy based on forced exploration

The non-adaptive ETC (Explore-then-Commit) approach outlined below is predicated on ex ante knowledge of the problem horizon $n$ (this is not a constraining factor since the exponential doubling trick [51] can be used to make the algorithm horizon-free) and a gap parameter $\delta \in(0, \Delta]$. In what follows, a new arm refers to one that is freshly queried from the arm-reservoir.

```
Algorithm 10 ALG1 \((n, \delta)\) (Non-adaptive ETC)
    Input: Horizon of play \(n\), gap parameter \(\delta\).
2: Set \(m=\left\lceil 2 \delta^{-2} \log n\right\rceil\).
3: New epoch: Play \(K=2\) new arms from the reservoir (call consideration set \(\mathcal{A}=\{1,2\}\) ).
4: Observe rewards \(\left(X_{1,1}, X_{2,1}\right)\).
    : Play each arm \(m-1\) times more; observe rewards \(\left(X_{1, j}, X_{2, j}: j=2, \ldots, m\right)\).
    if \(\left|\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right|<\delta m\) then
7: \(\quad\) Permanently discard \(\mathcal{A}\) and repeat from step (3).
    else
9: \(\quad\) Permanently commit to \(\operatorname{arm} i^{*} \in \arg \max _{i \in \mathcal{A}}\left\{\sum_{j=1}^{m} X_{i, j}\right\}\).
```

Policy dynamics. The horizon is divided into epochs of length $2 m=\Theta(\log n)$ each. In each epoch, the algorithm re-initializes by querying the arm-reservoir for a pair of new arms, and playing them $m$ times each. Subsequently, the pair is classified as either "distinct" or "identical"-typed via a hypothesis test (step 6 of Algorithm 10). If classified as distinct, the algorithm commits the residual budget of play to the empirically superior arm among the two (with ties broken arbitrarily). On the other hand, if the pair is classified as identical, the algorithm discards it permanently and ushers in a new epoch. The entire process repeats until a distinct-typed pair is identified. ALG1 $(n, \delta)$, albeit non-adaptive to $\Delta$, serves as an insightful basal motif for algorithm design and its operating
principle will guide the development of the $\Delta$-adaptive algorithm discussed next.

### 4.3.2 A UCB-based approach with adaptive resampling

In this section, we revisit the UCB-based policy in [21] for the static version of our problem (constant $\alpha(t)$ ). This policy requires ex ante knowledge of the support of reward distributions for parameter tuning. As a result, only "maximally supported" distributions such as Bernoulli(•), Uniform on $[0,1], \operatorname{Beta}(\cdot, \cdot)$, etc., are amenable to its performance guarantees; Uniform on $[0,0.5]$, on the other hand, is not. We fix this issue by adding a centered Gaussian noise term to the cumulative-difference-of-reward statistic in step (5) of Algorithm 11 to essentially create an unbounded support. This rids the algorithm of its dependence on assumptions pertaining to the support while preserving logarithmic regret in the statis setting. In fact, our modifications lead to a sharper characterization of the scaling of regret w.r.t. $\Delta$ in the static setting (see Theorem 16, 18).

Remark. The original policy (Algorithm 2 in cited reference) uses a threshold that is distinct from $4 \sqrt{m \log m}$ in step (5) of Algorithm 11; the choice of $4 \sqrt{m \log m}$ here aims to unify technical exposition and facilitate a fair comparison of upper bounds.

```
Algorithm 11 ALG2 (Nested UCB)
    New epoch \((t \leftarrow \mathbf{0})\) : Play \(K=2\) new arms from the reservoir (call set \(\mathcal{A}=\{1,2\}\) ).
    Observe rewards \(\left(X_{1,1}, X_{2,1}\right)\). Minimum per-arm sample count \(m \leftarrow 1\).
    Generate an independent sample of a standard Gaussian distribution \(\mathcal{Z}\).
    for \(t \in\{3,4, \ldots\}\) do
        if \(\left|\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right|<4 \sqrt{m \log m}\) then
            Permanently discard \(\mathcal{A}\) and repeat from step (1).
        else
            Play arm \(i_{t} \in \arg \max _{i \in \mathcal{A}}\left(\frac{\sum_{j=1}^{N_{i}(t-1)} X_{i, j}}{N_{i}(t-1)}+\sqrt{\frac{2 \log (t-1)}{N_{i}(t-1)}}\right)\).
            Observe reward \(X_{i_{t}, N_{i_{t}}(t)}\).
            if \(m<\min _{i \in \mathcal{A}} N_{i}(t)\) then
                        \(m \leftarrow m+1\).
```

Policy dynamics. ALG2 also has an episodic dynamic with exactly one pair of arms played per episode. It is noteworthy that ALG2 plays arms according to UCB1 [10] in every episode as opposed to playing them equally often until committing to the empirically superior one à la ALG1 $(n, \delta)$. Secondly, ALG2 never "commits" to an arm (or a consideration set); the implication is that the algorithm will keep querying new consideration sets from the reservoir throughout the horizon of play. Aforementioned adaptive resampling property is at the core of its horizon-free nature.

Operating principle. At any time, ALG2 computes a threshold of $O(\sqrt{m \log m})$ for the length- $m$ cumulative-difference-of-reward process, where $m$ denotes the minimum sample count among the two arms. If the envelope of said process is dominated by $O(\sqrt{m \log m})$, the arms are likely to belong to the same type (simultaneously optimal or inferior). The explanation stems from the Law of the Iterated Logarithm (see [25], Theorem 8.5.2): a zero-drift length- $m$ random walk process has its envelope bounded by $O(\sqrt{m \log \log m})$. In the aforementioned scenario, the algorithm discards the consideration set and subsequently, a new epoch is ushered in. This is done to avoid the possibility of incurring linear regret in the event that the two arms are inferior, since it is statistically impossible to distinguish a simultaneous-inferior consideration set from one where both arms are optimal, in the absence of any auxiliary information such as the mean rewards associated with the two types. On the other hand, if the cumulative-difference-of-reward dominates $O(\sqrt{m \log m})$, the consideration set is likely to contain arms of distinct types and ALG2 continues to run UCB1 on this set until the $O(\sqrt{m \log m})$ threshold is breached again.

Reason for introducing the Gaussian corruption. Centered Gaussian noise is added to the cumulative-difference-of-reward process in step (5) of ALG2 to avoid the possibility of incurring linear regret should the support of the reward distributions be a "very small" subset of $[0,1]$. To illustrate this point, suppose that the rewards associated with the types are deterministic with $\Delta<2 \sqrt{2 \log 2}$. Then, as soon as the algorithm queries a heterogeneous consideration set (one arm optimal and the other inferior) and the per-arm sample count reaches 2 , the cumulative-difference-of-reward statistic will satisfy $\left|\sum_{j=1}^{2}\left(X_{1, j}-X_{2, j}\right)\right|=2 \Delta<4 \sqrt{2 \log 2}$, resulting in the consideration
set getting discarded. On the other hand, if the consideration set is homogeneous (both arms optimal or inferior), the algorithm will still re-initialize within a finite number of samples in expectation (again, owing to the Law of the Iterated Logarithm). This will force the algorithm to keep querying new arms from the reservoir at rate that is linear in time, which is tantamount to incurring linear regret in the horizon. The addition of centered Gaussian noise hedges against this risk by guaranteeing that the cumulative-difference-of-reward process essentially has an infinite support at all times (even when the reward distributions might be degenerate). This rids the regret performance of its fragility w.r.t. ex ante knowledge of the support of reward distributions. The next proposition crystallizes this discussion.

Proposition 2 (Persistence of heterogeneous consideration sets) Let $\left\{X_{i, j}: j=1,2, \ldots\right\}$ be a collection of independent samples from an arm of type $i$ (implying $\mathbb{E}\left[X_{1, j}-X_{2, j}\right]=\Delta \forall j=1,2, .$. ). Let $\mathcal{Z}$ be an independently generated standard Gaussian random variable. Then,

$$
\begin{equation*}
\mathbb{P}\left(\bigcap_{m \geqslant 1}\left\{\left|\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{m \log m}\right\}\right)>\frac{\bar{\Phi}\left(f\left(T_{0}\right)\right)}{2}=: \beta_{\Delta}>0, \tag{4.2}
\end{equation*}
$$

where $\bar{\Phi}(v):=1 / \sqrt{2 \pi} \int_{v}^{\infty} e^{-u^{2} / 2} d u \forall v>0$ is the right-tail of the standard Gaussian CDF, $T_{0}:=$ $\left\lceil\left(64 / \Delta^{2}\right) \log ^{2}\left(64 / \Delta^{2}\right)\right\rceil$, and $f(x):=x+4 \sqrt{x \log x} \forall x \geqslant 1$.

Interpretation of $\boldsymbol{\beta}_{\Delta}$. First of all, note that $\beta_{\Delta}$ admits a closed-form characterization in terms of standard functions and satisfies $\beta_{\Delta}>0$ when $\Delta>0$ with $\lim _{\Delta \rightarrow 0} \beta_{\Delta}=0$. Secondly, $\beta_{\Delta}$ depends exclusively on $\Delta$, and represents a lower bound on the probability that ALG2 will never discard a consideration set containing arms of distinct types. This meta-result will be key to the upper bounds stated in forthcoming sections.

### 4.4 Natural models of the arm-reservoir

The probability $\alpha(t)$ of a newly queried arm at time $t$ being optimal-typed will likely vary over the horizon of play in realistic settings. For example, in the context of crowdsourcing applications,
the availability of "high quality" workers for a given task may depend on the prevailing populationlevel perception of the platform. This could plausibly be a function of the age $t$ of the platform. Alternatively, the reservoir may react to a query at time $t$ through its cumulative query count $\mathcal{J}_{t}$; this model may be suited to settings where, for example, acquiring a new resource is costly and yields diminishing returns. We capture these aspects through two reservoir models described next.

Model 1 (Exogenous reservoir) $(\alpha(t): t=1,2, \ldots)$ is a non-increasing deterministic process with $\alpha(1)=c \in(0,1)$, evolving independently of the decision maker's actions.

Model 2 (Endogenous reservoir) $(\alpha(t): t=1,2, \ldots)$ evolves as $\alpha(t)=g\left(\mathcal{J}_{t-1}\right)$, where $g: \mathbb{N} \cup$ $\{0\} \mapsto(0, c]$ is a non-increasing deterministic mapping with $g(0)=c \in(0,1)$, and $\mathcal{J}_{t}$ denotes the number of reservoir queries until time $t$ (inclusive) with $\mathcal{J}_{0}:=0$.

Our first result below states a necessary condition for achievability of sub-linear regret in the two reservoir models.

Theorem 14 (Necessary conditions for "complete learning" in the two reservoir models) 1 . Under Model 1, if $\sum_{t=1}^{\infty} \alpha(t)<\infty$, the expected cumulative regret of any policy $\pi$ grows linearly in the horizon of play, i.e., $\mathbb{E} R_{n}^{\pi}=\Omega(\Delta n)$, where the $\Omega(\cdot)$ only hides multiplicative constants independent of $\Delta$ and $n$. Equivalently, a necessary condition for achieving sublinear regret in the problem is $\sum_{t=1}^{\infty} \alpha(t)=\infty$.
2. Under Model 2, if $\sum_{t=1}^{\infty} g(t)<\infty$, the expected cumulative regret of any policy $\pi$ grows linearly in the horizon of play, i.e., $\mathbb{E} R_{n}^{\pi}=\Omega(\Delta n)$, where the $\Omega(\cdot)$ only hides multiplicative constants independent of $\Delta$ and $n$. Equivalently, a necessary condition for achieving sublinear regret in the problem is $\sum_{t=1}^{\infty} g(t)=\infty$.

The proof relies essentially on reduction to a full-information setting where the decision maker observes the true mean of an arm immediately upon play (refer to Appendix C for details). The optimal policy is this setting will keep querying the reservoir for new arms until it finds one with the
optimal mean, which it will subsequently commit the rest of its sampling budget to. The conditions stated in Theorem 14 are necessary for this policy to find an optimal arm within its lifetime. It is only natural that the same condition is necessary for achievability of sub-linear regret in the general setting where the decision maker only observes a noisy version of the true means. Surprisingly, however, these conditions are also almost-sufficient for sub-linear regret (as forthcoming results will show), and are therefore nearly-tight.

### 4.4.1 Exogenous reservoirs

The focus here will be on settings specified by Model 1. Theorem 14 establishes a necessary condition of $\sum_{t=1}^{\infty} \alpha(t)=\infty$ for achievability of sub-linear instance-dependent regret in the problem. In what follows, we show a slightly more refined characteristic: $\tilde{\Theta}\left(t^{-1}\right)$ is, in fact, a critical rate for "complete learning" in that policies achieving sub-linear regret exist if $\alpha(t)=\omega((\log t) / t)$.

To elucidate the criticality of the $\tilde{\Theta}\left(t^{-1}\right)$ rate, it will be convenient to consider a parameterization of $\alpha(t)$ given by $\alpha(t)=c t^{-\gamma}$, where $c$ is as specified in Model 1 and $\gamma \in[0,1)$ is a fixed parameter. This parameterization offers meaningful insights as to the statistical complexity of the problem w.r.t. $\gamma$ and facilitates an easy comparison between the regret guarantees of the algorithms discussed in §4.3. We will begin with an upper bound on the performance of Algorithm 10.

Theorem 15 (Upper bound for $\operatorname{ALG1}(n, \delta)$ ) Under Model 1 with $\alpha(t)=c t^{-\gamma}$, where $\gamma \in[0,1)$, the expected cumulative regret of the policy $\pi$ given by Algorithm 10 satisfies

$$
\limsup _{n \rightarrow \infty} \frac{\mathbb{E} R_{n}^{\pi}}{(\log n)^{\frac{1}{1-\gamma}}} \leqslant 24 \Delta\left(\frac{8}{\delta^{2} c}\right)^{\frac{1}{1-\gamma}} \mathfrak{F}\left(\left[\frac{\gamma}{1-\gamma}\right]\right),
$$

where $\mathfrak{F}(\cdot)$ denotes the Factorial function.

The proof of Theorem 15 involves technical details beyond the scope of a succinct discussion here (refer to Appendix C for details). The above result establishes that the inflation in regret as a result of asymptotically vanishing $\alpha(t)$ is at most poly-logarithmic in the horizon for "slowly
decaying" $\alpha(t)$, until a sharp phase transition to linear regret occurs around the $\alpha(t)=\tilde{\Theta}\left(t^{-1}\right)$ "critical rate" (see Theorem 14.1).

Remarks. (i) Logarithmic regret. Theorem 15 implies that $\mathbb{E} R_{n}^{\pi}=O\left((c \Delta)^{-1} \log n\right)$ when $\delta=\Delta$ and the reservoir has no "leakage" $(\gamma=0)$, consistent with known results for the static version of the problem [21]. (ii) Improving sample-efficiency. Instead of discarding both arms after step 6 of Algorithm 10, one can, in principle, discard only one arm, and query only one new arm from the reservoir as replacement. The regret incurred by this modified policy will differ only in absolute constants. The given design only intends to unify Algorithm 10 structurally with the other algorithm discussed in $\S 4.3$ so as to facilitate an easy comparison between the performance guarantees of the two algorithms.

Theorem 16 (Upper bound for ALG2) Under Model 1, the expected cumulative regret of the policy $\pi$ given by Algorithm 11 satisfies

$$
\limsup _{n \rightarrow \infty} \frac{\alpha(n) \mathbb{E} R_{n}^{\pi}}{\log n} \leqslant \frac{8}{\Delta \beta_{\Delta}},
$$

where $\beta_{\Delta}$ is as defined in (4.2).

The proof is provided in Appendix C.
Discussion. It follows directly from the above result that $\alpha(n)=\omega(\log n / n)$ is sufficient for ALG2 to be first-order optimal. On the other hand, we have already identified $\alpha(n)=\omega(1 / n)$ as a necessary condition for the existence of a first-order optimal policy (Theorem 14.1). Thus, the characterization of $t^{-1}$ as a critical rate in Theorem 14.1 is sharp up to a logarithmic scaling term.

Remark. The scaling factor $\beta_{\Delta}$ in Theorem 16 can, in fact, be shaved off entirely by introducing in ALG2 an initial "burn-in" phase (sub-linear and coercive in the horizon) during each epoch à la $\operatorname{ALG1}(n, \delta)$. This will, however, be achieved at the expense of ALG2's anytime property. Horizon-independence can then be restored by the use of a standard exponential doubling trick, see, e.g., [51]. The resulting algorithm is not discussed in this paper for brevity.

Comparison of theoretical performance. To facilitate a direct comparison between the two
upper bounds, it is conducive to consider $\alpha(n)=c n^{-\gamma}$ with $\gamma<1$ in Theorem 16. Evidently, ALG2 pays a heavy price for adaptivity to $\Delta$ which reflects in a polynomial $\tilde{O}\left(n^{\gamma}\right)$ regret as compared to the poly-logarithmic $O\left((\log n)^{\frac{1}{1-\gamma}}\right)$ regret achievable under $\operatorname{ALG1}(n, \delta)$. As to whether a performance gap between $\Delta$-aware and $\Delta$-adaptive algorithms is fundamental in Model 1 remains an open problem at the moment.

### 4.4.2 Endogenous reservoirs

The focus here will be on settings specified by Model 2 . We will begin with an upper bound on the theoretical performance of Algorithm 10. As before, it will be conducive to pivot to a parametric family of mappings $g(\cdot)$.

Theorem 17 (Upper bound for $\operatorname{ALG1}(n, \delta))$ Under Model 2 with $g(u)=c(u+1)^{-\gamma}$ for $u \geqslant 0$, where $\gamma \in[0,1)$ is a fixed parameter, the expected cumulative regret of the policy $\pi$ given by Algorithm 10 satisfies

$$
\limsup _{n \rightarrow \infty} \frac{\mathbb{E} R_{n}^{\pi}}{\log n} \leqslant\left(\frac{48 \Delta}{\delta^{2}}\right)\left(\frac{4}{c}\right)^{\frac{1}{1-\gamma}} \mathfrak{F}\left(\left\lceil\frac{\gamma}{1-\gamma}\right\rceil\right),
$$

where $\mathfrak{F}(\cdot)$ denotes the Factorial function.

The proof is provided in Appendix C. Evidently, unlike Theorem 15, only the multiplicative factors of the logarithmic term in Theorem 17 blow up as $\gamma$ approaches 1. Thus, under Model 2, a sharper phase transition from logarithmic to linear regret occurs at the critical reservoir-depletion rate of $\tilde{\Theta}\left(u^{-1}\right)$, where $u$ is the cumulative query count. We next look at the performance guarantee of ALG2 under Model 2.

Theorem 18 (Upper bound for ALG2) Under Model 2, the expected cumulative regret of the policy $\pi$ given by Algorithm 11 satisfies

$$
\limsup _{n \rightarrow \infty} \frac{\mathbb{E} R_{n}^{\pi}}{\log n} \leqslant\left(\frac{16 c}{\Delta}\right) \sum_{k=0}^{\infty} \exp \left(-\beta_{\Delta} \sum_{j=0}^{k-1} g(2 j)\right),
$$

where $\beta_{\Delta}$ is as defined in (4.2).

The proof is provided in Appendix C.
Remark. If the function $g(\cdot)$ is constant (equal to $c$ ), the upper bound in Theorem 18 is bounded above by $32 /\left(\Delta \beta_{\Delta}\right)$, which is independent of $c$. The implication is that achievable regret in the problem depends on the probability $c$ of sampling optimal-typed arms from the reservoir, surprisingly, only through sub-logarithmic terms when the reservoir is "static." This is fundamentally distinct from the upper bound in Theorem 17 which scales inversely with $c$ when $\gamma=0$.

Comparison of theoretical performance. Akin to Theorem 16, the $\beta_{\Delta}$ factor in Theorem 18 may also be shaved off, albeit at the expense of ALG2's anytime property. Then, for $g(u)=c(u+$ $1)^{-\gamma}$, the upper bound in Theorem 18 is bounded above by $(c / \Delta)(4 / c)^{1 /(1-\gamma)} \mathscr{F}(\lceil\gamma /(1-\gamma)\rceil)$. This order matches (up to numerical factors) the upper bound in Theorem 17. However, as $\gamma$ approaches 0 , the upper bound in Theorem 17 approaches a scaling of $c^{-1}$ while that in Theorem 18 is independent of $c$ in the limit. This observation suggests that there is merit to using ALG2 when the reservoir is static or nearly-static, i.e., when it suffers a negligible leakage (or loss) of optimal arms over time (or with increasing number of queries).

In a nutshell, the upper bounds in Theorem 17 and 18, combined with the lower bound in Theorem 14.2 , underscore the criticality of $g(u)=\tilde{\Theta}\left(u^{-1}\right)$ for achievability of sub-linear regret in the problem when there is attrition of optimal-typed arms from the reservoir with increasing number of queries.

### 4.5 Concluding remarks

This chapter summarizes the contributions in [68] and attempts to develop a principled approach to understanding the impact of endogenous variations in countably many-armed bandit problems through stylized arm-reservoir models. While we provide a sharp characterization of critical reservoir-decay rates necessary for achieving sub-linear regret in two natural arm-reservoir models, an important outstanding challenge is to identify more "reasonable" models of endogeneity and to design "robust" algorithms for such settings. To our best knowledge, the modeling
and methodological approaches adopted in this work are novel; we hope these may guide future research on some of the interesting directions emanating out of the results in this chapter.

## References

[1] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," Biometrika, vol. 25, no. 3/4, pp. 285-294, 1933.
[2] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multiarmed bandit problems," Machine Learning, vol. 5, no. 1, pp. 1-122, 2012.
[3] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," Advances in applied mathematics, vol. 6, no. 1, pp. 4-22, 1985.
[4] J.-Y. Audibert, R. Munos, and C. Szepesvári, "Exploration-exploitation tradeoff using variance estimates in multi-armed bandits," Theoretical Computer Science, vol. 410, no. 19, pp. 1876-1902, 2009.
[5] A. Garivier, T. Lattimore, and E. Kaufmann, "On explore-then-commit strategies," in Advances in Neural Information Processing Systems, 2016, pp. 784-792.
[6] A. Garivier and O. Cappé, "The kl-ucb algorithm for bounded stochastic bandits and beyond," in Proceedings of the 24th annual conference on learning theory, 2011, pp. 359376.
[7] S. Agrawal and N. Goyal, "Near-optimal regret bounds for thompson sampling," Journal of the ACM (JACM), vol. 64, no. 5, pp. 1-24, 2017.
[8] J.-Y. Audibert, S. Bubeck, et al., "Minimax policies for adversarial and stochastic bandits.," in COLT, vol. 7, 2009, pp. 1-122.
[9] T. Lattimore and C. Szepesvári, Bandit algorithms. Cambridge University Press, 2020.
[10] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," Machine learning, vol. 47, no. 2-3, pp. 235-256, 2002.
[11] J.-Y. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits.," in COLT, 2010, pp. 41-53.
[12] S. S. Villar, J. Bowden, and J. Wason, "Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges," Statistical science: a review journal of the Institute of Mathematical Statistics, vol. 30, no. 2, p. 199, 2015.
[13] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in Conference on learning theory, JMLR Workshop and Conference Proceedings, 2012, pp. 39-1.
[14] J. Shin, A. Ramdas, and A. Rinaldo, "Are sample means in multi-armed bandits positively or negatively biased?" Advances in Neural Information Processing Systems, vol. 32, 2019.
[15] -_, "On the bias, risk, and consistency of sample means in multi-armed bandits," SIAM Journal on Mathematics of Data Science, vol. 3, no. 4, pp. 1278-1300, 2021.
[16] -_, "On conditional versus marginal bias in multi-armed bandits," in International Conference on Machine Learning, PMLR, 2020, pp. 8852-8861.
[17] K. Zhang, L. Janson, and S. Murphy, "Inference for batched bandits," Advances in Neural Information Processing Systems, vol. 33, pp. 9818-9829, 2020.
[18] V. Hadad, D. A. Hirshberg, R. Zhan, S. Wager, and S. Athey, "Confidence intervals for policy evaluation in adaptive experiments," Proceedings of the National Academy of Sciences, vol. 118, no. 15, 2021.
[19] S. Wager and K. Xu, "Diffusion asymptotics for sequential experiments," arXiv preprint arXiv:2101.09855, 2021.
[20] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Computing Surveys (CSUR), vol. 54, no. 6, pp. 1-35, 2021.
[21] A. Kalvit and A. Zeevi, "From finite to countable-armed bandits," Advances in Neural Information Processing Systems, vol. 33, pp. 8259-8269, 2020.
[22] P. W. Glynn, "Diffusion approximations," Handbooks in Operations research and management Science, vol. 2, pp. 145-198, 1990.
[23] Y. Deshpande, L. Mackey, V. Syrgkanis, and M. Taddy, "Accurate inference for adaptive linear models," in International Conference on Machine Learning, PMLR, 2018, pp. 11941203.
[24] E. Kaufmann, O. Cappé, and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models," The Journal of Machine Learning Research, vol. 17, no. 1, pp. 1-42, 2016.
[25] R. Durrett, Probability: theory and examples. Cambridge university press, 2019, vol. 49.
[26] J. Honda and A. Takemura, "An asymptotically optimal bandit algorithm for bounded support models.," in COLT, Citeseer, 2010, pp. 67-79.
[27] M. Rothschild, "A two-armed bandit theory of market pricing," Journal of Economic Theory, vol. 9, no. 2, pp. 185-202, 1974.
[28] A. McLennan, "Price dispersion and incomplete learning in the long run," Journal of Economic dynamics and control, vol. 7, no. 3, pp. 331-347, 1984.
[29] L. Fan and P. W. Glynn, "Diffusion approximations for thompson sampling," arXiv preprint arXiv:2105.09232, 2021.
[30] C. Kalkanli and A. Ozgur, "Asymptotic convergence of thompson sampling," arXiv preprint arXiv:2011.03917, 2020.
[31] V. F. Araman and R. A. Caldentey, "Diffusion approximations for a class of sequential experimentation problems," Management Science, 2021.
[32] A. Kalvit and A. Zeevi, "A closer look at the worst-case behavior of multi-armed bandit algorithms," in Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 88078819.
[33] A. Slivkins et al., "Introduction to multi-armed bandits," Foundations and Trends® in Machine Learning, vol. 12, no. 1-2, pp. 1-286, 2019.
[34] S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi, "Mnl-bandit: A dynamic learning approach to assortment selection," Operations Research, vol. 67, no. 5, pp. 1453-1485, 2019.
[35] L. Bui, R. Johari, and S. Mannor, "Clustered bandits," arXiv preprint arXiv:1206.4169, 2012.
[36] S. Banerjee, S. Gollapudi, K. Kollias, and K. Munagala, "Segmenting two-sided markets," in Proceedings of the 26th International Conference on World Wide Web, 2017, pp. 63-72.
[37] R. Johari, V. Kamble, and Y. Kanoria, "Matching while learning," Operations Research, vol. 69, no. 2, pp. 655-681, 2021.
[38] D. A. Berry, R. W. Chen, A. Zame, D. C. Heath, L. A. Shepp, et al., "Bandit problems with infinitely many arms," The Annals of Statistics, vol. 25, no. 5, pp. 2103-2116, 1997.
[39] Y. Wang, J.-Y. Audibert, and R. Munos, "Algorithms for infinitely many-armed bandits," in Advances in Neural Information Processing Systems, 2009, pp. 1729-1736.
[40] T. Bonald and A. Proutiere, "Two-target algorithms for infinite-armed bandits with bernoulli rewards," in Advances in Neural Information Processing Systems, 2013, pp. 2184-2192.
[41] A. Carpentier and M. Valko, "Simple regret for infinitely many armed bandits," in International Conference on Machine Learning, 2015, pp. 1133-1141.
[42] H. P. Chan and S. Hu, "Infinite arms bandit: Optimality via confidence bounds," arXiv preprint arXiv:1805.11793, 2018.
[43] R. de Heide, J. Cheshire, P. Ménard, and A. Carpentier, "Bandits with many optimal arms," in Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 22 457-22 469.
[44] K. Chandrasekaran and R. Karp, "Finding a most biased coin with fewest flips," in Conference on Learning Theory, PMLR, 2014, pp. 394-407.
[45] K. Jamieson, D. Haas, and B. Recht, "On the detection of mixture distributions with applications to the most biased coin problem," arXiv preprint arXiv:1603.08037, 2016.
[46] A. R. Chaudhuri and S. Kalyanakrishnan, "Quantile-regret minimisation in infinitely manyarmed bandits.," in UAI, 2018, pp. 425-434.
[47] Y. Zhu and R. Nowak, "On regret with multiple best arms," Advances in Neural Information Processing Systems, vol. 33, pp. 9050-9060, 2020.
[48] A. Kalvit and A. Zeevi, "Complexity analysis of a countable-armed bandit problem," in International Conference on Algorithmic Learning Theory, 2023, ser. Proceedings of Machine Learning Research, vol. 201, PMLR, 2023, pp. 850-890.
[49] C. Derman, G. J. Lieberman, and S. M. Ross, "A sequential stochastic assignment problem," Management Science, vol. 18, no. 7, pp. 349-355, 1972.
[50] L. T. Liu, F. Ruan, H. Mania, and M. I. Jordan, "Bandit learning in decentralized matching markets," The Journal of Machine Learning Research, vol. 22, no. 1, pp. 9612-9645, 2021.
[51] L. Besson and E. Kaufmann, "What doubling tricks can and can't do for multi-armed bandits," arXiv preprint arXiv:1803.06971, 2018.
[52] A. Kalvit and A. Zeevi, "Dynamic learning in large matching markets," ACM SIGMETRICS Performance Evaluation Review, vol. 50, no. 2, pp. 18-20, 2022.
[53] J. Gittins and D. Jones, "A dynamic allocation index for the sequential design of experiments," Progress in statistics, J. Gani, Ed., pp. 241-266, 1974.
[54] P. Whittle, "Restless bandits: Activity allocation in a changing world," Journal of applied probability, pp. 287-298, 1988.
[55] R. R. Weber and G. Weiss, "On an index policy for restless bandits," Journal of applied probability, pp. 637-648, 1990.
[56] W. Hu and P. Frazier, "An asymptotically optimal index policy for finite-horizon restless bandits," arXiv preprint arXiv:1707.00205, 2017.
[57] J. Gittins, K. Glazebrook, and R. Weber, Multi-armed bandit allocation indices. John Wiley \& Sons, 2011.
[58] G. Zayas-Caban, S. Jasin, and G. Wang, "An asymptotically optimal heuristic for general nonstationary finite-horizon restless multi-armed, multi-action bandits," Advances in Applied Probability, vol. 51, no. 3, pp. 745-772, 2019.
[59] D. B. Brown and J. E. Smith, "Index policies and performance bounds for dynamic selection problems," Management Science, vol. 66, no. 7, pp. 3029-3050, 2020.
[60] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," SIAM journal on computing, vol. 32, no. 1, pp. 48-77, 2002.
[61] A. Slivkins and E. Upfal, "Adapting to a changing environment: The brownian restless bandits.," in COLT, 2008, pp. 343-354.
[62] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in International Conference on Algorithmic Learning Theory, Springer, 2011, pp. 174-188.
[63] O. Besbes, Y. Gur, and A. Zeevi, "Stochastic multi-armed-bandit problem with non-stationary rewards," Advances in neural information processing systems, vol. 27, pp. 199-207, 2014.
[64] N. Levine, K. Crammer, and S. Mannor, "Rotting bandits," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 3074-3083.
[65] R. Kleinberg and N. Immorlica, "Recharging bandits," in 2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2018, pp. 309-319.
[66] L. Cella and N. Cesa-Bianchi, "Stochastic bandits with delay-dependent payoffs," in International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 1168-1177.
[67] O. Besbes, Y. Gur, and A. Zeevi, "Optimal exploration-exploitation in a multi-armed bandit problem with non-stationary rewards," Stochastic Systems, vol. 9, no. 4, pp. 319-337, 2019.
[68] A. Kalvit and A. Zeevi, "Bandits with dynamic arm-acquisition costs," in 2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2022, pp. 1-7.
[69] W Hoeffding, "Probability inequalities for sums of bounded random variables," Journal of the American Statistical Association, vol. 58, no. 301, pp. 13-30, 1963.
[70] P. Billingsley, Convergence of probability measures. John Wiley \& Sons, 2013.
[71] P. Flajolet, D. Gardy, and L. Thimonier, "Birthday paradox, coupon collectors, caching algorithms and self-organizing search," Discrete Applied Mathematics, vol. 39, no. 3, pp. 207229, 1992.

## Appendix A: Appendix to Chapter 1

## General organization

1. §A. 1 provides closed-form expressions for the $\lambda_{\rho}^{*}(\theta)$ and $h_{\rho}(\theta)$ functions that appear in Theorem 1 and Theorem 4.
2. §A. 2 states three ancillary results that will be used in other proofs.
3. §A. 3 provides the proof of Theorem 1 in the "large gap" regime.
4. §A. 4 provides the proof of Theorem 1 in the "small gap" regime.
5. §A. 5 provides the proof of Theorem 1 in the "moderate gap" regime.
6. §A. 6 provides the proof of Theorem 3.
7. §A. 7 provides the proof of Theorem 4.
8. §A. 8 provides the proof of Theorem 5 .
9. §A. 9 provides the proof of Theorem 2.
10. §A. 10 provides proofs for the ancillary results stated in Appendix A.2.

Additional notation. In the proofs that follow, $\Gamma \cdot\rceil$ has been used to denote the "ceiling operator," i.e., $\lceil x\rceil=\inf \{v \in \mathbb{N}: v \geqslant x\}$ for any $x \in \mathbb{R}$. Similarly, $\lfloor\cdot\rfloor$ denotes the "floor operator," i.e., $\lfloor x\rfloor=\sup \{v \in \mathbb{N}: v \leqslant x\}$ for any $x \in \mathbb{R}$.

## A. 1 Closed-form expressions for $\lambda_{\rho}^{*}(\theta)$ and $h_{\rho}(\theta)$

$\lambda_{\rho}^{*}(\theta)$ is given by:

$$
\begin{equation*}
\lambda_{\rho}^{*}(\theta)=\frac{1}{2}+\sqrt{\frac{1}{4}-\frac{1}{\left(1+\sqrt{1+\frac{\theta}{\rho}}\right)^{2}}} . \tag{A.1}
\end{equation*}
$$

$h_{\rho}(\theta)$ is given by:

$$
\begin{equation*}
h_{\rho}(\theta)=\sqrt{\theta}\left(\frac{1}{2}-\sqrt{\frac{1}{4}-\frac{1}{\left(1+\sqrt{1+\frac{\theta}{\rho}}\right)^{2}}}\right) . \tag{A.2}
\end{equation*}
$$

## A. 2 Auxiliary results

We will use the following version of the Chernoff-Hoeffding inequality [69] in our proofs:

Fact 1 Chernoff-Hoeffding bound. Suppose that $\left\{Y_{i, j}: i \in\{1,2\}, j \in \mathbb{N}\right\}$ is a collection of independent, zero-mean random variables such that $\forall i \in\{1,2\}, j \in \mathbb{N}, Y_{i, j} \in\left[c_{i}, 1+c_{i}\right]$ almost surely, for some fixed $c_{1}, c_{2} \leqslant 0$. Then, for any $m_{1}, m_{2} \in \mathbb{N}$ and $\alpha>0$,

$$
\mathbb{P}\left(\frac{\sum_{j=1}^{m_{1}} Y_{1, j}}{m_{1}}-\frac{\sum_{j=1}^{m_{2}} Y_{2, j}}{m_{2}} \geqslant \alpha\right) \leqslant \exp \left(\frac{-2 \alpha^{2} m_{1} m_{2}}{m_{1}+m_{2}}\right) .
$$

Proof. Let $[n]:=\{1, \ldots, n\}$ for $n \in \mathbb{N}$. The Chernoff-Hoeffding inequality in its standard form states that for independent, zero-mean, bounded random variables $\left\{Z_{j}: j \in[n]\right\}$ with $Z_{j} \in$ $\left[a_{j}, b_{j}\right] \forall j \in[n]$, the following holds for any $\varepsilon>0$,

$$
\begin{equation*}
\mathbb{P}\left(\sum_{j=1}^{n} Z_{j} \geqslant \varepsilon n\right) \leqslant \exp \left(\frac{-2 \varepsilon^{2} n^{2}}{\sum_{j=1}^{n}\left(b_{j}-a_{j}\right)^{2}}\right) \tag{A.3}
\end{equation*}
$$

The desired form of the inequality can be obtained by making the following substitutions in (A.3): $n \leftarrow m_{1}+m_{2} ; Z_{j} \leftarrow \frac{Y_{1, j}}{m_{1}}, a_{j} \leftarrow \frac{c_{1}}{m_{1}}, b_{j} \leftarrow \frac{1+c_{1}}{m_{1}}$ for $j \in\left[m_{1}\right] ; Z_{j} \leftarrow \frac{-Y_{2, j-m_{1}}}{m_{2}}, a_{j} \leftarrow \frac{-\left(1+c_{2}\right)}{m_{2}}$,
$b_{j} \leftarrow \frac{-c_{2}}{m_{2}}$ for $j \in\left[m_{1}+m_{2}\right] \backslash\left[m_{1}\right]$; and $\varepsilon \leftarrow \frac{\alpha}{m_{1}+m_{2}}$, in that order.

In addition, we will use in the proof of Theorem 3 the following two properties of the Beta distribution:

Fact 2 If $\theta_{k}, \tilde{\theta}_{k}$ are Beta $(1, k+1)$-distributed, and $\theta_{k}, \tilde{\theta}_{l}$ are independent $\forall k, l \in \mathbb{N} \cup\{0\}$, then

$$
\mathbb{P}\left(\theta_{k}>\tilde{\theta}_{l}\right)=\frac{l+1}{k+l+2} \quad \text { for any } k, l \in \mathbb{N} \cup\{0\}
$$

Fact 3 If $\theta_{k}, \tilde{\theta}_{k}$ are $\operatorname{Beta}(k+1,1)$-distributed, and $\theta_{k}, \tilde{\theta}_{l}$ are independent $\forall k, l \in \mathbb{N} \cup\{0\}$, then

$$
\mathbb{P}\left(\theta_{k}>\tilde{\theta}_{l}\right)=\frac{k+1}{k+l+2} \quad \text { for any } k, l \in \mathbb{N} \cup\{0\}
$$

The proofs of Fact (2) and Fact (3) are elementary and relegated to Appendix A. 10.

## A. 3 Proof of Theorem 1 in the "large gap" regime

The proof is straightforward in this regime. We know that for $\rho>1, \mathbb{E} R_{n}^{\pi} \leqslant C \rho\left(\frac{\log n}{\Delta}+\frac{\Delta}{\rho-1}\right)$ for some absolute constant $C$ (see [4], Theorem 7). Since $\mathbb{E} R_{n}^{\pi}=\Delta \mathbb{E}\left[n-N_{i^{*}}(n)\right]$, it follows that $\mathbb{E}\left[\frac{n-N_{i^{*}}(n)}{n}\right]=o(1)$ in the "large gap" regime. Using Markov's inequality, we then conclude that $\frac{n-N_{i^{*}}(n)}{n}=o_{p}(1)$, or equivalently, $\lim _{n \rightarrow \infty} \frac{N_{i^{*}}(n)}{n}=1$. Results for "small" and "moderate" gaps are provided separately in Appendix A. 4 and Appendix A. 5 respectively.

## A. 4 Proof of Theorem 1 in the "small gap" regime

Without loss of generality, suppose that arm 1 is optimal, i.e., $\mu_{1} \geqslant \mu_{2}$. We will show that for any $\epsilon>0$, it follows that $\lim _{n \rightarrow \infty} \mathbb{P}\left(\frac{N_{1}(n)}{n} \geqslant \frac{1}{2}+\epsilon\right)=0$. Then, since arm 2 is inferior, an identical result would naturally hold for it as well. Combining the two would prove our assertion as desired. To this end, pick an arbitrary $\epsilon \in(0,1 / 2)$, define $u(n):=\left\lceil\left(\frac{1}{2}+\epsilon\right) n\right\rceil$, and consider the following:

$$
\begin{align*}
N_{1}(n) & \leqslant u(n)+\sum_{t=u(n)+1}^{n} \mathbb{1}\left\{\pi_{t}=1, N_{1}(t-1) \geqslant u(n)\right\} \\
& =u(n)+\sum_{t=u(n)}^{n-1} \mathbb{1}\left\{\pi_{t+1}=1, N_{1}(t) \geqslant u(n)\right\} \\
& \leqslant u(n)+\sum_{t=u(n)}^{n-1} \mathbb{1}\left\{\pi_{t+1}=1, N_{1}(t) \geqslant u(t)\right\} \\
& \leqslant u(n)+\sum_{t=u(n)}^{n-1} \mathbb{1}\left\{\bar{X}_{1}(t)-\bar{X}_{2}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{2}(t)}}-\frac{1}{\sqrt{N_{1}(t)}}\right), N_{1}(t) \geqslant u(t)\right\} \\
& =u(n)+\underbrace{\sum_{t=u(n)}^{n-1} \mathbb{1}\left\{\bar{Y}_{1}(t)-\bar{Y}_{2}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{2}(t)}}-\frac{1}{\sqrt{N_{1}(t)}}\right)-\Delta, N_{1}(t) \geqslant u(t)\right\}}_{=: Z(n)}, \tag{A.4}
\end{align*}
$$

where $\bar{Y}_{i}(t):=\frac{\sum_{j=1}^{N_{i}(t)} Y_{i, j}}{N_{i}(t)}$ with $Y_{i, j}:=X_{i, j}-\mu_{i}, i \in\{1,2\}, j \in \mathbb{N}$. Clearly, $Y_{i, j}$ 's are independent, zero-mean, and $Y_{i, j} \in\left[-\mu_{i}, 1-\mu_{i}\right] \forall i \in\{1,2\}, j \in \mathbb{N}$.

## A.4.1 An almost sure lower bound on the arm-sampling rates

As a meta-result, we will first show that $N_{i}(n) / n$, for both arms $i \in\{1,2\}$, is bounded away from 0 by a positive constant, almost surely. To this end, consider $n$ large enough such that for the $\epsilon$ selected earlier, we have $\Delta<\sqrt{\frac{\rho \log n}{n}}\left(\frac{1}{\sqrt{1 / 2-\epsilon}}-\frac{1}{\sqrt{1 / 2+\epsilon}}\right)$; this is possible since $\Delta=o\left(\sqrt{\frac{\log n}{n}}\right)$ in the "small gap" regime. Working with a large enough $n$ will allow us to use the Chernoff-Hoeffding bound (Fact 1) in step ( $\star$ ) in the forthcoming analysis. Observe from (A.4) that

$$
\begin{align*}
& \mathbb{E} Z(n) \\
&= \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{1}(t)-\bar{Y}_{2}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{2}(t)}}-\frac{1}{\sqrt{N_{1}(t)}}\right)-\Delta, N_{1}(t) \geqslant u(t)\right) \\
&=\sum_{t=u(n)}^{n-1} \sum_{m=u(t)}^{t-1} \mathbb{P}\left(\frac{\sum_{j=1}^{m} Y_{1, j}}{m}-\frac{\sum_{j=1}^{t-m} Y_{2, j}}{t-m} \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{t-m}}-\frac{1}{\sqrt{m}}\right)-\Delta, N_{1}(t)=m\right) \\
& \leqslant \sum_{(\star)}^{n-1} \sum_{t=u(n)}^{t-1} \mathbb{P}\left(\frac{\sum_{j=1}^{m} Y_{1, j}}{m}-\frac{\sum_{j=1}^{t-m} Y_{2, j}}{t-m} \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{t-m}}-\frac{1}{\sqrt{m}}\right)-\Delta\right) \\
& \leqslant\left.\left.\sum_{(\dagger)}^{n-1} \sum_{t=u(n)}^{t-1} \exp \left[-2 \rho\left(1-2 \sqrt{\frac{m}{t}\left(1-\frac{m}{t}\right.}\right)\right) \log t+4 \Delta \sqrt{\rho t \log t}\left(\sqrt{\frac{m}{t}}-\sqrt{1-\frac{m}{t}}\right) \sqrt{\frac{m}{t}\left(1-\frac{m}{t}\right.}\right)\right] \\
& \leqslant \sum_{(\ddagger)}^{n-1} \sum_{t=u(n)}^{t-1} \exp \left[-2 \rho\left(1-\sqrt{1-4 \epsilon^{2}}\right) \log t\right] \exp \left[4 \Delta \sqrt{\rho t \log t}\left(\sqrt{\frac{m}{t}}-\sqrt{1-\frac{m}{t}}\right) \sqrt{\frac{m}{t}\left(1-\frac{m}{t}\right)}\right] \\
& \leqslant \sum_{t=u(n)}^{n-1} \sum_{m=u(t)}^{t-1} \exp \left[-2 \rho\left(1-\sqrt{1-4 \epsilon^{2}}\right) \log t\right] \exp [4 \Delta \sqrt{\rho t \log t}] \\
& \leqslant \sum_{t=u(n)}^{n-1} \sum_{m=u(t)}^{t-1} \exp \left[-2 \rho\left(1-\sqrt{1-4 \epsilon^{2}}\right) \log t\right] \exp [4 \Delta \sqrt{\rho n \log n}] \\
& \leqslant \exp [4 \Delta \sqrt{\rho n \log n}] \sum_{t=u(n)}^{n-1} t^{-\left(2 \rho-1-2 \rho \sqrt{1-4 \epsilon^{2}}\right)} \\
& \leqslant \exp [o(4 \sqrt{\rho} \log n)] \sum_{t=u(n)}^{n-1} t^{-\left(2 \rho-1-2 \rho \sqrt{1-4 \epsilon^{2}}\right)} \\
& \leqslant n^{\frac{1}{2}-\epsilon} \sum_{t=u(n)}^{n-1} t^{-\left(2 \rho-1-2 \rho \sqrt{1-4 \epsilon^{2}}\right)}, \tag{A.5}
\end{align*}
$$

where $(\dagger)$ follows after an application of the Chernoff-Hoeffding bound (Fact 1$)$, ( $\ddagger$ ) since $\frac{m}{t}\left(1-\frac{m}{t}\right) \leqslant$ $\frac{1}{4}-\epsilon^{2}$ on the interval $\{m: u(t) \leqslant m \leqslant t-1\}$, and the last inequality in (A.5) holds for $n$ large
enough. Now consider an arbitrary $\delta>0$. Then,

$$
\begin{array}{rlr}
\mathbb{P}\left(N_{1}(n)-u(n) \geqslant \delta n\right) & \leqslant \mathbb{P}(Z(n) \geqslant \delta n) & \quad \text { (using (A.4)) } \\
& \leqslant \frac{\mathbb{E} Z(n)}{\delta n} & \leqslant\left(\frac{n^{-\left(\frac{1}{2}+\epsilon\right)}}{\delta}\right) \sum_{t=u(n)}^{n-1} t^{-\left(2 \rho-1-2 \rho \sqrt{1-4 \epsilon^{2}}\right)} \\
\Longrightarrow \mathbb{P}\left(\frac{N_{1}(n)}{n} \geqslant \frac{1}{2}+\epsilon+\delta+\frac{1}{n}\right) & \leqslant\left(\frac{n^{-\left(\frac{1}{2}+\epsilon\right)}}{\delta}\right) \sum_{t=\left\lceil\frac{n}{2}\right\rceil}^{n-1} t^{-\left(2 \rho-1-2 \rho \sqrt{1-4 \epsilon^{2}}\right)} .
\end{array}
$$

Define $g(\rho, \epsilon):=\frac{1}{2}+\epsilon+2 \rho-1-2 \rho \sqrt{1-4 \epsilon^{2}}$. Since $\rho>1$ is fixed, and $\epsilon \in(0,1 / 2)$ is arbitrary, it is possible to push $\epsilon$ close to $1 / 2$ to ensure that $g(\rho, \epsilon)>2$. Therefore, $\exists \epsilon_{\rho} \in(0,1 / 2)$ s.t. $g(\rho, \epsilon)>2$ for $\epsilon \geqslant \epsilon_{\rho}$. Plugging in $\epsilon=\epsilon_{\rho}$ in (A.6), we obtain

$$
\mathbb{P}\left(\frac{N_{1}(n)}{n} \geqslant \frac{1}{2}+\epsilon_{\rho}+\delta+\frac{1}{n}\right) \leqslant\left(\frac{2^{2 \rho-1}}{\delta}\right) n^{-\left(g\left(\rho, \epsilon_{\rho}\right)-1\right)} .
$$

Note that since $\epsilon_{\rho}<1 / 2, \exists \epsilon_{\rho}^{\prime}<1 / 2$ s.t. $\epsilon_{\rho}+1 / n<\epsilon_{\rho}^{\prime}$ for $n$ large enough, i.e., the following holds for all $n$ large enough:

$$
\mathbb{P}\left(\frac{N_{1}(n)}{n} \geqslant \frac{1}{2}+\epsilon_{\rho}^{\prime}+\delta\right) \leqslant\left(\frac{2^{2 \rho-1}}{\delta}\right) n^{-\left(g\left(\rho, \epsilon_{\rho}\right)-1\right)} .
$$

Finally, since $\delta>0$ is arbitrary, and $g\left(\rho, \epsilon_{\rho}\right)>2$, it follows from the Borel-Cantelli Lemma that

$$
\limsup _{n \rightarrow \infty} \frac{N_{1}(n)}{n} \leqslant \frac{1}{2}+\epsilon_{\rho}^{\prime}<1 \quad \text { w.p. } 1 .
$$

By assumption, arm 2 is inferior; the above result thus holds, in fact, for both the arms (An almost identical proof can be replicated for rigor). Therefore, we conclude

$$
\begin{equation*}
\liminf _{n \rightarrow \infty} \frac{N_{i}(n)}{n} \geqslant \frac{1}{2}-\epsilon_{\rho}^{\prime}>0 \quad \text { w.p. } 1 \quad \forall i \in\{1,2\} . \tag{A.7}
\end{equation*}
$$

## A.4.2 Closing the loop

In this part of the proof, we will leverage (A.7) to finally show that $N_{i}(n) / n=1 / 2+o_{p}(1)$ for $i \in\{1,2\}$. To this end, recall from (A.4) that

$$
\begin{align*}
& \mathbb{E} Z(n) \\
= & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{1}(t)-\bar{Y}_{2}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{2}(t)}}-\frac{1}{\sqrt{N_{1}(t)}}\right)-\Delta, N_{1}(t) \geqslant u(t)\right) \\
\leqslant & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{1}(t)-\bar{Y}_{2}(t) \geqslant \sqrt{\frac{\rho \log t}{t}}\left(\frac{1}{\sqrt{\frac{1}{2}-\epsilon}}-\frac{1}{\sqrt{\frac{1}{2}+\epsilon}}\right)-\Delta, N_{1}(t) \geqslant u(t)\right) \\
\leqslant & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{1}(t)-\bar{Y}_{2}(t) \geqslant \sqrt{\frac{\rho \log t}{t}}\left(\frac{1}{\sqrt{\frac{1}{2}-\epsilon}}-\frac{1}{\sqrt{\frac{1}{2}+\epsilon}}\right)-\Delta\right) \\
= & \sum_{t=u(n)}^{n-1} \mathbb{P}(\underbrace{\sqrt{\frac{t}{\rho \log t}}\left(\bar{Y}_{1}(t)-\bar{Y}_{2}(t)\right)}_{=: W_{t}} \geqslant \sqrt{\frac{2}{1-2 \epsilon}}-\sqrt{\frac{2}{1+2 \epsilon}}-\Delta \sqrt{\frac{t}{\rho \log t}}) \\
\leqslant & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(W_{t} \geqslant \frac{1}{\sqrt{1-2 \epsilon}}-\frac{1}{\sqrt{1+2 \epsilon}}\right), \tag{A.8}
\end{align*}
$$

for $n$ large enough; the last inequality following since $\Delta=o\left(\sqrt{\frac{\log n}{n}}\right)$ and $u(n)>n / 2$. Now,

$$
\begin{align*}
& \left|W_{t}\right| \\
\leqslant & \sqrt{\frac{t}{\rho \log t}}\left(\left|\frac{\sum_{j=1}^{N_{1}(t)} Y_{1, j}}{N_{1}(t)}\right|+\left|\frac{\sum_{j=1}^{N_{2}(t)} Y_{2, j}}{N_{2}(t)}\right|\right) \\
= & \sqrt{\frac{2 t}{\rho \log t}}\left(\left.\sqrt{\frac{\log \log N_{1}(t)}{N_{1}(t)}}\left|\frac{\sum_{j=1}^{N_{1}(t)} Y_{1, j}}{\sqrt{2 N_{1}(t) \log \log N_{1}(t)}}\right|+\sqrt{\frac{\log \log N_{2}(t)}{N_{2}(t)}} \right\rvert\, \frac{\sum_{j=1}^{N_{2}(t)} Y_{2, j}}{\sqrt{2 N_{2}(t) \log \log N_{2}(t)}}\right. \\
\leqslant & \sqrt{\frac{2 t}{\rho \log t}}\left(\sqrt{\frac{\log \log t}{N_{1}(t)}}\left|\frac{\sum_{j=1}^{N_{1}(t)} Y_{1, j}}{\sqrt{2 N_{1}(t) \log \log N_{1}(t)}}\right|+\sqrt{\frac{\log \log t}{N_{2}(t)}}\left|\frac{\sum_{j=1}^{N_{2}(t)} Y_{2, j}}{\sqrt{2 N_{2}(t) \log \log N_{2}(t)}}\right|\right) \\
= & \sqrt{\frac{2 \log \log t}{\rho \log t}}\left(\sqrt{\frac{t}{N_{1}(t)}}\left|\frac{\sum_{j=1}^{N_{1}(t)} Y_{1, j}}{\sqrt{2 N_{1}(t) \log \log N_{1}(t)}}\right|+\sqrt{\frac{t}{N_{2}(t)}}\left|\frac{\sum_{j=1}^{N_{2}(t)} Y_{2, j}}{\sqrt{2 N_{2}(t) \log \log N_{2}(t)}}\right|\right) . \tag{A.9}
\end{align*}
$$

We know that $N_{i}(t)$, for both arms $i \in\{1,2\}$, can be lower bounded path-wise by a deterministic monotone increasing function of $t$, say $f(t)$, that grows to $+\infty$ as $t \rightarrow \infty$. This is a trivial consequence of the structure of canonical UCB (Algorithm 1), and the fact that the rewards are uniformly bounded. We therefore have for any $i \in\{1,2\}$ that

$$
\left|\frac{\sum_{j=1}^{N_{i}(t)} Y_{i, j}}{\sqrt{2 N_{i}(t) \log \log N_{i}(t)}}\right| \leqslant \sup _{m \geqslant f(t)}\left|\frac{\sum_{j=1}^{m} Y_{i, j}}{\sqrt{2 m \log \log m}}\right| .
$$

For a fixed arm $i \in\{1,2\},\left\{Y_{i, j}: j \in \mathbb{N}\right\}$ is a collection of i.i.d. random variables with $\mathbb{E} Y_{i, 1}=0$ and $\operatorname{Var}\left(Y_{i, 1}\right)=\operatorname{Var}\left(X_{i, 1}\right) \leqslant 1$. Also, $f(t)$ is monotone increasing and coercive in $t$. Therefore, the Law of the Iterated Logarithm (see [25], Theorem 8.5.2) implies

$$
\begin{equation*}
\limsup _{t \rightarrow \infty}\left|\frac{\sum_{j=1}^{N_{i}(t)} Y_{i, j}}{\sqrt{2 N_{i}(t) \log \log N_{i}(t)}}\right| \leqslant 1 \quad \text { w.p. } 1 \quad \forall i \in\{1,2\} \text {. } \tag{A.10}
\end{equation*}
$$

Using (A.7), (A.9) and (A.10), we conclude that

$$
\begin{equation*}
\lim _{t \rightarrow \infty} W_{t}=0 \quad \text { w.p. } 1 . \tag{A.11}
\end{equation*}
$$

Now consider an arbitrary $\delta>0$. Then,

$$
\begin{align*}
\mathbb{P}\left(N_{1}(n)-u(n) \geqslant \delta n\right) & \leqslant \mathbb{P}(Z(n) \geqslant \delta n)  \tag{A.4}\\
& \leqslant \frac{\mathbb{E} Z(n)}{\delta n} \\
& \leqslant \frac{1}{\delta n} \sum_{t=u(n)}^{n-1} \mathbb{P}\left(W_{t} \geqslant \frac{1}{\sqrt{1-2 \epsilon}}-\frac{1}{\sqrt{1+2 \epsilon}}\right) .  \tag{A.8}\\
& \leqslant \frac{1}{\delta} \sup _{t>n / 2} \mathbb{P}\left(W_{t} \geqslant \frac{1}{\sqrt{1-2 \epsilon}}-\frac{1}{\sqrt{1+2 \epsilon}}\right) \\
\Longrightarrow \mathbb{P}\left(\frac{N_{1}(n)}{n} \geqslant \frac{1}{2}+\epsilon+\delta+\frac{1}{n}\right) & \leqslant \frac{1}{\delta} \sup _{t>n / 2} \mathbb{P}\left(W_{t} \geqslant \frac{1}{\sqrt{1-2 \epsilon}}-\frac{1}{\sqrt{1+2 \epsilon}}\right) .
\end{align*}
$$

(Markov's inequality)

Since $\epsilon, \delta>0$ are arbitrary, it follows that for $n$ large enough,

$$
\begin{equation*}
\mathbb{P}\left(\frac{N_{1}(n)}{n} \geqslant \frac{1}{2}+2(\epsilon+\delta)\right) \leqslant \frac{1}{\delta} \sup _{t>n / 2} \mathbb{P}\left(W_{t} \geqslant \frac{1}{\sqrt{1-2 \epsilon}}-\frac{1}{\sqrt{1+2 \epsilon}}\right) . \tag{A.12}
\end{equation*}
$$

Using (A.11) and (A.12), we conclude that for any arbitrary $\epsilon, \delta>0$,

$$
\limsup _{n \rightarrow \infty} \mathbb{P}\left(\frac{N_{1}(n)}{n} \geqslant \frac{1}{2}+2(\epsilon+\delta)\right) \leqslant \frac{1}{\delta} \limsup _{n \rightarrow \infty} \mathbb{P}\left(W_{n} \geqslant \frac{1}{\sqrt{1-2 \epsilon}}-\frac{1}{\sqrt{1+2 \epsilon}}\right)=0 .
$$

It therefore follows that for any $\epsilon^{\prime}>0, \lim _{n \rightarrow \infty} \mathbb{P}\left(\frac{N_{1}(n)}{n} \geqslant \frac{1}{2}+\epsilon^{\prime}\right)=0$; equivalently, we have $\lim _{n \rightarrow \infty} \mathbb{P}\left(\frac{N_{2}(n)}{n} \leqslant \frac{1}{2}-\epsilon^{\prime}\right)=0$. Since arm 2 is inferior by assumption, it naturally holds that $\lim _{n \rightarrow \infty} \mathbb{P}\left(\frac{N_{2}(n)}{n} \geqslant \frac{1}{2}+\epsilon^{\prime}\right)=0$ (The steps in A.4.2 can be replicated near-identically for rigor). Thus, the stated assertion $\frac{N_{i}(n)}{n} \xrightarrow[n \rightarrow \infty]{p} \frac{1}{2} \forall i \in\{1,2\}$, follows.

## A. 5 Proof of Theorem 1 in the "moderate gap" regime

Firstly, note that the $\lambda_{\rho}^{*}(\theta)$ that solves (1.2), satisfies the following properties: (i) Continuous and monotone increasing in $\theta \geqslant 0$, (ii) $\lambda_{\rho}^{*}(\theta) \geqslant 1 / 2$ for all $\theta \geqslant 0$, (iii) $\lambda_{\rho}^{*}(0)=1 / 2$ and $\lambda_{\rho}^{*}(\theta) \rightarrow 1$ as $\theta \rightarrow \infty$.

Secondly, because we are only interested in asymptotics, the $\Delta \sim \sqrt{\frac{\theta \log n}{n}}$ condition is as good as $\Delta=\sqrt{\frac{\theta \log n}{n}}$, since for any arbitrarily small $\epsilon^{\prime}>0, \Delta \in\left(\sqrt{\frac{\left(\theta-\epsilon^{\prime}\right) \log n}{n}}, \sqrt{\frac{\left(\theta+\epsilon^{\prime}\right) \log n}{n}}\right)$ for $n$ large enough; the stated assertion would follow in the limit as $\epsilon^{\prime}$ approaches 0 . In what follows, we will therefore assume for readability of the proof, and without loss of generality, that $\Delta=\sqrt{\frac{\theta \log n}{n}}$.

Thirdly, without loss of generality, suppose that arm 1 is optimal, i.e., $\mu_{1} \geqslant \mu_{2}$.

## A.5.1 Focusing on arm 1

Consider an arbitrary $\epsilon \in\left(0,1-\lambda_{\rho}^{*}(\theta)\right)$, and define $u(n):=\left\lceil\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right) n\right\rceil$. We know that

$$
\begin{align*}
N_{1}(n) & \leqslant u(n)+\sum_{t=u(n)+1}^{n} \mathbb{1}\left\{\pi_{t}=1, N_{1}(t-1) \geqslant u(n)\right\} \\
& =u(n)+\sum_{t=u(n)}^{n-1} \mathbb{1}\left\{\pi_{t+1}=1, N_{1}(t) \geqslant u(n)\right\} \\
& \leqslant u(n)+\sum_{t=u(n)}^{n-1} \mathbb{1}\left\{\pi_{t+1}=1, N_{1}(t) \geqslant u(t)\right\} \\
& \leqslant u(n)+\sum_{t=u(n)}^{n-1} \mathbb{1}\left\{\bar{X}_{1}(t)-\bar{X}_{2}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{2}(t)}}-\frac{1}{\sqrt{N_{1}(t)}}\right), N_{1}(t) \geqslant u(t)\right\} \\
& =u(n)+\underbrace{\sum_{t=u(n)}^{n-1} \mathbb{1}\left\{\bar{Y}_{1}(t)-\bar{Y}_{2}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{2}(t)}}-\frac{1}{\sqrt{N_{1}(t)}}\right)-\Delta, N_{1}(t) \geqslant u(t)\right\}}_{=: Z(n)} \tag{A.13}
\end{align*}
$$

where $\bar{Y}_{i}(t):=\frac{\sum_{j=1}^{N_{i}(t)} Y_{i, j}}{N_{i}(t)}$ with $Y_{i, j}:=X_{i, j}-\mu_{i}, i \in\{1,2\}, j \in \mathbb{N}$. Clearly, $Y_{i, j}$ 's are independent, zero-mean, and $Y_{i, j} \in\left[-\mu_{i}, 1-\mu_{i}\right] \forall i \in\{1,2\}, j \in \mathbb{N}$.

## An almost sure lower bound on the arm-sampling rates

Consider $n$ large enough such that $\sqrt{\frac{\log n}{n}}$ is monotone decreasing in $n$ ( $n \geqslant 3$ suffices). This will enable the inequality in step $(\dagger)$ below. From (A.13), we have

$$
\begin{align*}
& \mathbb{E} Z(n) \\
= & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{1}(t)-\bar{Y}_{2}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{2}(t)}}-\frac{1}{\sqrt{N_{1}(t)}}\right)-\Delta, N_{1}(t) \geqslant u(t)\right) \\
= & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{1}(t)-\bar{Y}_{2}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{2}(t)}}-\frac{1}{\sqrt{N_{1}(t)}}\right)-\sqrt{\frac{\theta \log n}{n}}, N_{1}(t) \geqslant u(t)\right) \\
\leqslant & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{1}(t)-\bar{Y}_{2}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{2}(t)}}-\frac{1}{\sqrt{N_{1}(t)}}\right)-\sqrt{\frac{\theta \log t}{t}}, N_{1}(t) \geqslant u(t)\right) \\
= & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{1}(t)-\bar{Y}_{2}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{2}(t)}}-\frac{1}{\sqrt{N_{1}(t)}}-\sqrt{\frac{\theta}{\rho t}}\right), N_{1}(t) \geqslant u(t)\right)  \tag{A.14}\\
\leqslant & \sum_{t=u(n)}^{n-1} \sum_{m=u(t)}^{t-1} \mathbb{P}\left(\frac{\sum_{j=1}^{m} Y_{1, j}}{m}-\frac{\sum_{j=1}^{t-m} Y_{2, j}}{t-m} \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{t-m}}-\frac{1}{\sqrt{m}}-\sqrt{\frac{\theta}{\rho t}}\right)\right) . \tag{A.15}
\end{align*}
$$

Notice that in the interval $m \in[u(t), t-1]$,

$$
\begin{aligned}
\frac{1}{\sqrt{t-m}}-\frac{1}{\sqrt{m}}-\sqrt{\frac{\theta}{2 t} \geqslant \frac{1}{\sqrt{t-u(t)}}-\frac{1}{\sqrt{u(t)}}-\sqrt{\frac{\theta}{\rho t}} \geqslant \frac{1}{\sqrt{t}}\left(\frac{1}{\sqrt{1-\lambda_{\rho}^{*}(\theta)-\epsilon}}-\frac{1}{\sqrt{\lambda_{\rho}^{*}(\theta)+\epsilon}}-\sqrt{\frac{\theta}{\rho}}\right)} \\
>0,
\end{aligned}
$$

where the final inequality follows since $\lambda_{\rho}^{*}(\theta)$ is the solution to (1.2). We can therefore apply the Chernoff-Hoeffding bound (Fact 1) to (A.15) to conclude

$$
\begin{align*}
& \mathbb{E} Z(n) \\
\leqslant & \sum_{t=u(n)}^{n-1} \sum_{m=u(t)}^{t-1} \exp \left[-2 \rho \log t\left(\frac{1}{\sqrt{t-m}}-\frac{1}{\sqrt{m}}-\sqrt{\frac{\theta}{\rho t}}\right)^{2} \frac{m(t-m)}{t}\right] \\
= & \sum_{t=u(n)}^{n-1} \sum_{m=u(t)}^{t-1} \exp \left[-2 \rho \log t\left(\sqrt{\frac{m}{t}}-\sqrt{1-\frac{m}{t}}-\sqrt{\frac{\theta}{\rho}} \sqrt{\frac{m}{t}\left(1-\frac{m}{t}\right)}\right)^{2}\right] \\
= & \sum_{t=u(n)}^{n-1} \sum_{m=u(t)}^{t-1} \exp \left[-2 \rho \log t\left(f\left(\frac{m}{t}\right)\right)^{2}\right], \tag{A.16}
\end{align*}
$$

where the function $f(x):=\sqrt{x}-\sqrt{1-x}-\sqrt{\theta x(1-x) / \rho}$. Notice that $f(x)$ is monotone increasing over the interval $(1 / 2,1)(\because \theta, \rho \geqslant 0)$. Also, note that $1 / 2<\lambda_{\rho}^{*}(\theta)+\epsilon<m / t<1$ in (A.16). Thus, we have in (A.16) that $f\left(\frac{m}{t}\right) \geqslant \min _{x \in\left[\lambda_{\rho}^{*}(\theta)+\epsilon, 1\right)} f(x)=f\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right)$. An expression for $f\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right)$ is provided in (A.19) below. Observe that $f\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right)>0$; this follows since $\lambda_{\rho}^{*}(\theta)$ is the solution to (1.2). Using these facts in (A.16), we conclude

$$
\begin{align*}
\mathbb{E} Z(n) & \leqslant \sum_{t=u(n)}^{n-1} \sum_{m=u(t)}^{t-1} \exp \left[-2 \rho \log t\left(\min _{x \in\left[\lambda_{\rho}^{*}(\theta)+\epsilon, 1\right)} f(x)\right)^{2}\right] \\
& =\sum_{t=u(n)}^{n-1} \sum_{m=u(t)}^{t-1} \exp \left[-2 \rho \log t\left(f\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right)\right)^{2}\right] \\
& \leqslant \sum_{t=u(n)}^{n-1} t^{1-2 \rho\left(f\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right)\right)^{2}} \\
& =\sum_{t=\left\lceil\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right) n\right\rceil}^{n-1} t^{1-2 \rho\left(f\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right)\right)^{2}} . \tag{A.17}
\end{align*}
$$

Now consider an arbitrary $\delta>0$. We then have

$$
\begin{align*}
\mathbb{P}\left(N_{1}(n)-u(n) \geqslant \delta n\right) & \leqslant \mathbb{P}(Z(n) \geqslant \delta n)  \tag{A.13}\\
& \leqslant \frac{\mathbb{E} Z(n)}{\delta n} \\
\Longrightarrow \mathbb{P}\left(N_{1}(n) \geqslant\left\lceil\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right) n\right\rceil+\delta n\right) & \leqslant \frac{1}{\delta n} \sum_{t=\left\lceil\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right) n\right\rceil}^{n-1} t^{1-2 \rho\left(f\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right)\right)^{2}} . \tag{A.17}
\end{align*}
$$

(Markov's inequality)

Note that $f\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right)$ is given by

$$
\begin{equation*}
f\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right)=\sqrt{\lambda_{\rho}^{*}(\theta)+\epsilon}-\sqrt{1-\lambda_{\rho}^{*}(\theta)-\epsilon}-\sqrt{\frac{\theta}{\rho}} \sqrt{\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right)\left(1-\lambda_{\rho}^{*}(\theta)-\epsilon\right)} . \tag{A.19}
\end{equation*}
$$

Setting $\epsilon=0$ in (A.19) yields $f\left(\lambda_{\rho}^{*}(\theta)\right)=0$ (follows from (1.2)), whereas setting $\epsilon=1-\lambda_{\rho}^{*}(\theta)$ yields $f(1)=1$. Since $\rho>1$, and $f\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right)$ is continuous and monotone increasing in $\epsilon$, $\exists \epsilon_{\theta, \rho} \in\left(0,1-\lambda_{\rho}^{*}(\theta)\right)$ s.t. $f\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right)>1 / \sqrt{\rho}$ for $\epsilon \geqslant \epsilon_{\theta, \rho}$. Substituting $\epsilon=\epsilon_{\theta, \rho}$ in (A.18) and using the aforementioned fact, we obtain

$$
\begin{align*}
\mathbb{P}\left(N_{1}(n) \geqslant\left\lceil\left(\lambda_{\rho}^{*}(\theta)+\epsilon_{\theta, \rho}\right) n\right\rceil+\delta n\right) & \leqslant \frac{1}{\delta n} \sum_{t=\left\lceil\left(\lambda_{\rho}^{*}(\theta)+\epsilon_{\theta, \rho}\right) n\right\rceil}^{n-1} t^{1-2 \rho\left(f\left(\lambda_{\rho}^{*}(\theta)+\epsilon_{\theta, \rho}\right)\right)^{2}} \\
& \leqslant\left(\frac{2^{2 \rho-1}}{\delta}\right) n^{-\left(2 \rho\left(f\left(\lambda_{\rho}^{*}(\theta)+\epsilon_{\theta, \rho}\right)\right)^{2}-1\right)} \tag{A.20}
\end{align*}
$$

where the last inequality follows since $1 / \sqrt{\rho}<f\left(\lambda_{\rho}^{*}(\theta)+\epsilon_{\theta, \rho}\right)<1$, and $\lambda_{\rho}^{*}(\theta)+\epsilon_{\theta, \rho}>1 / 2$. Finally since $\delta>0$ is arbitrary, we conclude from (A.20) using the Borel-Cantelli Lemma that

$$
\limsup _{n \rightarrow \infty} \frac{N_{1}(n)}{n} \leqslant \lambda_{\rho}^{*}(\theta)+\epsilon_{\theta, \rho}<1 \quad \text { w.p. } 1 .
$$

The above result naturally holds for arm 2 as well, since it is inferior by assumption (we resort to
the cop-out that a near-identical argument handles its case). Therefore, in conclusion,

$$
\begin{equation*}
\liminf _{n \rightarrow \infty} \frac{N_{i}(n)}{n} \geqslant 1-\lambda_{\rho}^{*}(\theta)-\epsilon_{\theta, \rho}>0 \quad \text { w.p. } 1 \quad \forall i \in\{1,2\} . \tag{A.21}
\end{equation*}
$$

## Closing the loop

From (A.14), we know that

$$
\begin{align*}
& \mathbb{E} Z(n) \\
\leqslant & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{1}(t)-\bar{Y}_{2}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{2}(t)}}-\frac{1}{\sqrt{N_{1}(t)}}-\sqrt{\frac{\theta}{\rho t}}\right), N_{1}(t) \geqslant u(t)\right) \\
\leqslant & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{1}(t)-\bar{Y}_{2}(t) \geqslant \sqrt{\frac{\rho \log t}{t}}\left(\frac{1}{\sqrt{1-\lambda_{\rho}^{*}(\theta)-\epsilon}}-\frac{1}{\sqrt{\lambda_{\rho}^{*}(\theta)+\epsilon}}-\sqrt{\frac{\theta}{\rho}}\right)\right) \\
= & \sum_{t=u(n)}^{n-1} \mathbb{P}(\underbrace{\sqrt{\frac{t}{\rho \log t}}\left(\bar{Y}_{1}(t)-\bar{Y}_{2}(t)\right)}_{=: W_{t}} \geqslant \frac{1}{\sqrt{1-\lambda_{\rho}^{*}(\theta)-\epsilon}}-\frac{1}{\sqrt{\lambda_{\rho}^{*}(\theta)+\epsilon}}-\sqrt{\frac{\theta}{\rho}}), \tag{A.22}
\end{align*}
$$

where we already know that $\frac{1}{\sqrt{1-\lambda_{\rho}^{*}(\theta)-\epsilon}}-\frac{1}{\sqrt{\lambda_{\rho}^{*}(\theta)+\epsilon}}-\sqrt{\frac{\theta}{\rho}}>0\left(\right.$ since $\lambda_{\rho}^{*}(\theta)$ is the solution to (1.2)). Now,

$$
\begin{align*}
& \left|W_{t}\right| \\
\leqslant & \sqrt{\frac{t}{\rho \log t}}\left(\left|\frac{\sum_{j=1}^{N_{1}(t)} Y_{1, j}}{N_{1}(t)}\right|+\left|\frac{\sum_{j=1}^{N_{2}(t)} Y_{2, j}}{N_{2}(t)}\right|\right) \\
= & \sqrt{\frac{2 t}{\rho \log t}}\left(\left.\sqrt{\frac{\log \log N_{1}(t)}{N_{1}(t)}}\left|\frac{\sum_{j=1}^{N_{1}(t)} Y_{1, j}}{\sqrt{2 N_{1}(t) \log \log N_{1}(t)}}\right|+\sqrt{\frac{\log \log N_{2}(t)}{N_{2}(t)}} \right\rvert\, \frac{\sum_{j=1}^{N_{2}(t)} Y_{2, j}}{\sqrt{2 N_{2}(t) \log \log N_{2}(t)}}\right. \\
\leqslant & \sqrt{\frac{2 t}{\rho \log t}}\left(\sqrt{\frac{\log \log t}{N_{1}(t)}}\left|\frac{\sum_{j=1}^{N_{1}(t)} Y_{1, j}}{\sqrt{2 N_{1}(t) \log \log N_{1}(t)}}\right|+\sqrt{\frac{\log \log t}{N_{2}(t)}}\left|\frac{\sum_{j=1}^{N_{2}(t)} Y_{2, j}}{\sqrt{2 N_{2}(t) \log \log N_{2}(t)}}\right|\right) \\
= & \sqrt{\frac{2 \log \log t}{\rho \log t}}\left(\sqrt{\frac{t}{N_{1}(t)}}\left|\frac{\sum_{j=1}^{N_{1}(t)} Y_{1, j}}{\sqrt{2 N_{1}(t) \log \log N_{1}(t)}}\right|+\sqrt{\frac{t}{N_{2}(t)}}\left|\frac{\sum_{j=1}^{N_{2}(t)} Y_{2, j}}{\sqrt{2 N_{2}(t) \log \log N_{2}(t)}}\right|\right) . \quad \text { (A.23) } \tag{A.23}
\end{align*}
$$

We know that $N_{i}(t)$, for both arms $i \in\{1,2\}$, can be lower bounded path-wise by a deterministic monotone increasing function of $t$, say $g(t)$, that grows to $+\infty$ as $t \rightarrow \infty$. This is a trivial consequence of the structure of the canonical UCB policy (Algorithm 1), and the fact that the rewards are uniformly bounded. Therefore, for any arm $i \in\{1,2\}$, we have

$$
\left|\frac{\sum_{j=1}^{N_{i}(t)} Y_{i, j}}{\sqrt{2 N_{i}(t) \log \log N_{i}(t)}}\right| \leqslant \sup _{m \geqslant g(t)}\left|\frac{\sum_{j=1}^{m} Y_{i, j}}{\sqrt{2 m \log \log m}}\right| .
$$

For a fixed $i \in\{1,2\},\left\{Y_{i, j}: j \in \mathbb{N}\right\}$ is a collection of i.i.d. random variables with $\mathbb{E} Y_{i, 1}=0$ and $\operatorname{Var}\left(Y_{i, 1}\right)=\operatorname{Var}\left(X_{i, 1}\right) \leqslant 1$. Also, $g(t)$ is a monotone increasing and coercive function of $t$. Therefore, the Law of the Iterated Logarithm (see [25], Theorem 8.5.2) implies

$$
\begin{equation*}
\limsup _{t \rightarrow \infty}\left|\frac{\sum_{j=1}^{N_{i}(t)} Y_{i, j}}{\sqrt{2 N_{i}(t) \log \log N_{i}(t)}}\right| \leqslant 1 \quad \text { w.p. } 1 \forall i \in\{1,2\} \text {. } \tag{A.24}
\end{equation*}
$$

Using (A.21), (A.23) and (A.24), we conclude that

$$
\begin{equation*}
\lim _{t \rightarrow \infty} W_{t}=0 \quad \text { w.p. } 1 \tag{A.25}
\end{equation*}
$$

Now consider an arbitrary $\delta>0$. We have

$$
\begin{align*}
& \mathbb{P}\left(N_{1}(n)-u(n) \geqslant \delta n\right) \\
\leqslant & \mathbb{P}(Z(n) \geqslant \delta n)  \tag{A.13}\\
\leqslant & \frac{\mathbb{E} Z(n)}{\delta n} \\
\leqslant & \frac{1}{\delta n} \sum_{t=u(n)}^{n-1} \mathbb{P}\left(W_{t} \geqslant \frac{1}{\sqrt{1-\lambda_{\rho}^{*}(\theta)-\epsilon}}-\frac{1}{\sqrt{\lambda_{\rho}^{*}(\theta)+\epsilon}}-\sqrt{\frac{\theta}{\rho}}\right)  \tag{A.22}\\
\leqslant & \frac{1}{\delta} \sup _{t>n / 2} \mathbb{P}\left(W_{t} \geqslant \frac{1}{\sqrt{1-\lambda_{\rho}^{*}(\theta)-\epsilon}}-\frac{1}{\sqrt{\lambda_{\rho}^{*}(\theta)+\epsilon}}-\sqrt{\frac{\theta}{\rho}}\right) . \tag{A.26}
\end{align*}
$$

Using (A.25) and (A.26), it follows that

$$
\limsup _{n \rightarrow \infty} \mathbb{P}\left(N_{1}(n)-u(n) \geqslant \delta n\right) \leqslant \frac{1}{\delta} \limsup _{n \rightarrow \infty} \mathbb{P}\left(W_{n} \geqslant \frac{1}{\sqrt{1-\lambda_{\rho}^{*}(\theta)-\epsilon}}-\frac{1}{\sqrt{\lambda_{\rho}^{*}(\theta)+\epsilon}}-\sqrt{\frac{\theta}{\rho}}\right)=0 .
$$

Since $u(n)=\left\lceil\left(\lambda_{\rho}^{*}(\theta)+\epsilon\right) n\right\rceil$ and $\epsilon, \delta>0$ are arbitrary, we conclude that for any $\epsilon>0$, it holds that $\lim _{n \rightarrow \infty} \mathbb{P}\left(\frac{N_{1}(n)}{n} \geqslant \lambda_{\rho}^{*}(\theta)+\epsilon\right)=0$. Equivalently, $\lim _{n \rightarrow \infty} \mathbb{P}\left(\frac{N_{2}(n)}{n} \leqslant 1-\lambda_{\rho}^{*}(\theta)-\epsilon\right)=0$ holds for any $\epsilon>0$.

## A.5.2 Focusing on arm 2 and concluding

We will essentially replicate here the proof for arm 1 given in A.5.1, albeit with a few subtle modifications to account for the fact that arm 2 is inferior. Consistent with previous approach and notation, we consider an arbitrary $\epsilon \in\left(0, \lambda_{\rho}^{*}(\theta)\right)$ and set $u(n):=\left\lceil\left(1-\lambda_{\rho}^{*}(\theta)+\epsilon\right) n\right\rceil$, where $\lambda_{\rho}^{*}(\theta)$ is the solution to (1.2) (Note that the definition of $u(n)$ here is different from the one used in the proof for arm 1.). We know that

$$
\begin{align*}
N_{2}(n) & \leqslant u(n)+\sum_{t=u(n)+1}^{n} \mathbb{1}\left\{\pi_{t}=2, N_{2}(t-1) \geqslant u(n)\right\} \\
& =u(n)+\sum_{t=u(n)}^{n-1} \mathbb{1}\left\{\pi_{t+1}=2, N_{2}(t) \geqslant u(n)\right\} \\
& \leqslant u(n)+\sum_{t=u(n)}^{n-1} \mathbb{1}\left\{\bar{X}_{2}(t)-\bar{X}_{1}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{1}(t)}}-\frac{1}{\sqrt{N_{2}(t)}}\right), N_{2}(t) \geqslant u(n)\right\} \\
& =u(n)+\underbrace{\sum_{t=u(n)}^{n-1} \mathbb{1}\left\{\bar{Y}_{2}(t)-\bar{Y}_{1}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{1}(t)}}-\frac{1}{\sqrt{N_{2}(t)}}\right)+\Delta, N_{2}(t) \geqslant u(n)\right\},}_{=: Z(n)} \tag{A.27}
\end{align*}
$$

where $\bar{Y}_{i}(t):=\frac{\sum_{j=1}^{N_{i}(t)} Y_{i, j}}{N_{i}(t)}$ with $Y_{i, j}:=X_{i, j}-\mu_{i}, i \in\{1,2\}, j \in \mathbb{N}$ (Notice that these definitions of $\bar{Y}_{i}(t)$ and $Y_{i, j}$ are identical to their counterparts from the proof for arm 1.). From (A.27), it follows
that

$$
\begin{aligned}
& \mathbb{E} Z(n) \\
= & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{2}(t)-\bar{Y}_{1}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{1}(t)}}-\frac{1}{\sqrt{N_{2}(t)}}\right)+\Delta, N_{2}(t) \geqslant u(n)\right) \\
= & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{2}(t)-\bar{Y}_{1}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{1}(t)}}-\frac{1}{\sqrt{N_{2}(t)}}\right)+\sqrt{\frac{\theta \log n}{n}}, N_{2}(t) \geqslant u(n)\right) \\
\leqslant & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{2}(t)-\bar{Y}_{1}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{t-u(n)}}-\frac{1}{\sqrt{u(n)}}\right)+\sqrt{\frac{\theta \log n}{n}}\right) \\
\leqslant & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{2}(t)-\bar{Y}_{1}(t) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{n-u(n)}}-\frac{1}{\sqrt{u(n)}}\right)+\sqrt{\frac{\theta \log t}{n}}\right) \\
\leqslant & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{2}(t)-\bar{Y}_{1}(t) \geqslant \sqrt{\frac{\rho \log t}{n}}\left(\frac{1}{\sqrt{\lambda_{\rho}^{*}(\theta)-\epsilon}}-\frac{1}{\sqrt{1-\lambda_{\rho}^{*}(\theta)+\epsilon}}+\sqrt{\frac{\theta}{\rho}}\right)\right),
\end{aligned}
$$

where $\frac{1}{\sqrt{\lambda_{\rho}^{*}(\theta)-\epsilon}}-\frac{1}{\sqrt{1-\lambda_{\rho}^{*}(\theta)+\epsilon}}+\sqrt{\frac{\theta}{\rho}}>0$ is guaranteed since $\lambda_{\rho}^{*}(\theta)$ is the solution to (1.2). Also, $t \geqslant u(n)=\left\lceil\left(1-\lambda_{\rho}^{*}(\theta)+\epsilon\right) n\right\rceil \Longrightarrow n \leqslant \frac{t}{1-\lambda_{\rho}^{*}(\theta)+\epsilon}$. Therefore,

$$
\begin{align*}
& \mathbb{E} Z(n) \\
\leqslant & \sum_{t=u(n)}^{n-1} \mathbb{P}\left(\bar{Y}_{2}(t)-\bar{Y}_{1}(t) \geqslant \sqrt{1-\lambda_{\rho}^{*}(\theta)+\epsilon} \sqrt{\frac{\rho \log t}{t}}\left(\frac{1}{\sqrt{\lambda_{\rho}^{*}(\theta)-\epsilon}}-\frac{1}{\sqrt{1-\lambda_{\rho}^{*}(\theta)+\epsilon}}+\sqrt{\frac{\theta}{\rho}}\right)\right) \\
= & \sum_{t=u(n)}^{n-1} \mathbb{P}(\underbrace{\sqrt{\frac{t}{\rho \log t}}\left(\bar{Y}_{2}(t)-\bar{Y}_{1}(t)\right)}_{=: W_{t}} \geqslant \underbrace{\sqrt{1-\lambda_{\rho}^{*}(\theta)+\epsilon}\left(\frac{1}{\sqrt{\lambda_{\rho}^{*}(\theta)-\epsilon}}-\frac{1}{\sqrt{1-\lambda_{\rho}^{*}(\theta)+\epsilon}}+\sqrt{\frac{\theta}{\rho}}\right)}_{=: \varepsilon(\theta, \rho, \epsilon)}) . \tag{A.28}
\end{align*}
$$

Recall that we have already handled $W_{t}$ (albeit a negated version thereof) in the proof for arm 1 in (A.22) and shown that $W_{t} \rightarrow 0$ almost surely in (A.25). Now consider an arbitrary $\delta>0$. We
then have

$$
\begin{align*}
\mathbb{P}\left(N_{2}(n)-u(n) \geqslant \delta n\right) & \leqslant \mathbb{P}(Z(n) \geqslant \delta n)  \tag{A.27}\\
& \leqslant \frac{\mathbb{E} Z(n)}{\delta n} \\
& \leqslant \frac{1}{\delta n} \sum_{t=u(n)}^{n-1} \mathbb{P}\left(W_{t} \geqslant \varepsilon(\theta, \rho, \epsilon)\right)  \tag{A.28}\\
& \leqslant \frac{1}{\delta} \sup _{t>\left(1-\lambda_{\rho}^{*}(\theta)\right) n} \mathbb{P}\left(W_{t} \geqslant \varepsilon(\theta, \rho, \epsilon)\right) . \tag{A.29}
\end{align*}
$$

(Markov's inequality)

Taking limits on both sides of (A.29), we obtain

$$
\limsup _{n \rightarrow \infty} \mathbb{P}\left(N_{2}(n)-u(n) \geqslant \delta n\right) \leqslant \frac{1}{\delta} \underset{n \rightarrow \infty}{\limsup } \mathbb{P}\left(W_{n} \geqslant \varepsilon(\theta, \rho, \epsilon)\right)=0
$$

where the final conclusion follows since $W_{n} \rightarrow 0$ almost surely, and hence also in probability. Now since $u(n)=\left\lceil\left(1-\lambda_{\rho}^{*}(\theta)+\epsilon\right) n\right\rceil$ and $\epsilon, \delta>0$ are arbitrary, it follows that for any $\epsilon>0$, we have $\lim _{n \rightarrow \infty} \mathbb{P}\left(\frac{N_{2}(n)}{n} \geqslant 1-\lambda_{\rho}^{*}(\theta)+\epsilon\right)=0$. From the proof for arm 1, we already know that $\lim _{n \rightarrow \infty} \mathbb{P}\left(\frac{N_{2}(n)}{n} \leqslant 1-\lambda_{\rho}^{*}(\theta)-\epsilon\right)=0$ holds for any $\epsilon>0$. Therefore, it must be the case that $\frac{N_{2}(n)}{n} \xrightarrow[n \rightarrow \infty]{p} 1-\lambda_{\rho}^{*}(\theta)$ and $\frac{N_{1}(n)}{n} \xrightarrow[n \rightarrow \infty]{p} \lambda_{\rho}^{*}(\theta)$, as desired.

## A. 6 Proof of Theorem 3

## A.6.1 Proof of part (I)

Let $\theta_{k}, \tilde{\theta}_{k}$ be $\operatorname{Beta}(1, k+1)$-distributed, with $\theta_{k}, \tilde{\theta}_{l}$ independent $\forall k, l \in \mathbb{N} \cup\{0\}$. In the twoarmed bandit with deterministic 0 rewards, at any time $n+1$, the probability of playing arm 1 conditioned on the entire history up to that point, is given by $\mathbb{P}\left(\pi_{n+1}=1 \mid \mathcal{F}_{n}\right)=\mathbb{P}\left(\theta_{N_{1}(n)}>\tilde{\theta}_{N_{2}(n)} \mid \mathcal{F}_{n}\right)=$ $\frac{n-N_{1}(n)+1}{n+2}$ (using Fact (2)). Since the arms are identical, and $N_{1}(n)+N_{2}(n)=n$, we must have $\mathbb{E}\left(N_{1}(n) / n\right)=1 / 2 \forall n \in \mathbb{N}$ by symmetry. Define $Z_{n}:=N_{1}(n) / n$. Then, $Z_{n}$ evolves according to
the following Markovian rule:

$$
Z_{n+1}=\left(\frac{n}{n+1}\right) Z_{n}+\frac{Y\left(Z_{n}, n, \xi_{n}\right)}{n+1}
$$

where $\left\{\xi_{n}\right\}$ is an independent noise process that is such that $Y\left(Z_{n}, n, \xi_{n}\right) \mid Z_{n}$ is distributed as Bernoulli $\left(\frac{n\left(1-Z_{n}\right)+1}{n+2}\right)$. Note that $Y(\cdot, \cdot, \cdot) \in\{0,1\}$. Then,

$$
\begin{aligned}
Z_{n+1}^{2} & =\left(\frac{n}{n+1}\right)^{2} Z_{n}^{2}+\left(\frac{Y\left(Z_{n}, n, \xi_{n}\right)}{n+1}\right)^{2}+\frac{2 n Z_{n} Y\left(Z_{n}, n, \xi_{n}\right)}{(n+1)^{2}} \\
& =\left(\frac{n}{n+1}\right)^{2} Z_{n}^{2}+\frac{Y\left(Z_{n}, n, \xi_{n}\right)}{(n+1)^{2}}+\frac{2 n Z_{n} Y\left(Z_{n}, n, \xi_{n}\right)}{(n+1)^{2}}
\end{aligned}
$$

Solving the recursion for $Z_{n+1}^{2}$, we obtain

$$
Z_{n+1}^{2}=\frac{Z_{1}^{2}+\sum_{t=1}^{n}\left[Y\left(Z_{t}, t, \xi_{t}\right)+2 t Z_{t} Y\left(Z_{t}, t, \xi_{t}\right)\right]}{(n+1)^{2}}=\frac{Z_{1}+\sum_{t=1}^{n}\left[Y\left(Z_{t}, t, \xi_{t}\right)+2 t Z_{t} Y\left(Z_{t}, t, \xi_{t}\right)\right]}{(n+1)^{2}},
$$

where the last equality follows since $Z_{1} \in\{0,1\}$. Taking expectations and using the fact that $\mathbb{E} Z_{t}=1 / 2 \forall t \in \mathbb{N}$, yields

$$
\mathbb{E} Z_{n+1}^{2}=\frac{\frac{1}{2}+\sum_{t=1}^{n}\left(\frac{1}{2}+2 t \mathbb{E}\left[\frac{t Z_{t}\left(1-Z_{t}\right)+Z_{t}}{t+2}\right]\right)}{(n+1)^{2}}
$$

Using $Z_{t}\left(1-Z_{t}\right) \leqslant 1 / 4$, we get the relation

$$
\mathbb{E} Z_{n+1}^{2} \leqslant \frac{1+\sum_{t=1}^{n}\left(1+t \mathbb{E}\left[\frac{t+4 Z_{t}}{t+2}\right]\right)}{2(n+1)^{2}}=\frac{n+1+\sum_{t=1}^{n} t}{2(n+1)^{2}}=\frac{n+2}{4(n+1)} .
$$

Thus, $\operatorname{Var}\left(\frac{N_{1}(n)}{n}\right)=\operatorname{Var}\left(Z_{n}\right) \leqslant \frac{n+1}{4 n}-\frac{1}{4}=\frac{1}{4 n}$. Since $\mathbb{E}\left(\frac{N_{1}(n)}{n}\right)=\frac{1}{2}$, we conclude using Chebyshev's inequality that $\frac{N_{1}(n)}{n} \rightarrow \frac{1}{2}$ in probability as $n \rightarrow \infty$.

## A.6.2 Proof of part (II)

Our proof of this part is essentially pivoted on showing the stronger result that $\mathbb{P}\left(N_{1}(n)=m\right)=$ $\frac{1}{n+1}$ for any $m \in\{0, \ldots, n\}$ and $n \in \mathbb{N}$. To this end, for an arbitrary $m$ in said interval, let $\mathrm{S}_{m}$ be the set of sample-paths of length $n$ such that $N_{1}(n)=m$ on each sample-path $\mathrm{s}_{m} \in \mathrm{~S}_{m}$. Clearly, $\left|\mathrm{S}_{m}\right|=\binom{n}{m}$. Let $i\left(\mathrm{~s}_{m}, t\right) \in\{1,2\}$ denote the index of the arm pulled at time $t \in\{1, \ldots, n\}$ on $\mathrm{s}_{m}$, and let $\tilde{N}_{j}(t)$ denote the number of pulls of arm $j \in\{1,2\}$ up to (and including) time $t$ on $\mathrm{s}_{m}$ (with $\left.\tilde{N}_{1}(0)=\tilde{N}_{2}(0):=0\right)$. Note that $i\left(\mathrm{~s}_{m}, t\right), \tilde{N}_{1}(t)$ and $\tilde{N}_{2}(t)$ are deterministic for all $t \in\{1, \ldots, n\}$, once $\mathrm{s}_{m}$ is fixed. Let $\theta_{k}, \tilde{\theta}_{k}$ be $\operatorname{Beta}(k+1,1)$-distributed, with $\theta_{k}, \tilde{\theta}_{l}$ independent $\forall k, l \in \mathbb{N} \cup\{0\}$. It then follows that

$$
\begin{align*}
\mathbb{P}\left(N_{1}(n)=m\right) & =\sum_{\mathrm{s}_{m} \in \mathrm{~S}_{m}} \prod_{t=1}^{n} \mathbb{P}\left(\theta_{\tilde{N}_{i\left(s_{m}, t\right)}(t-1)}>\tilde{\theta}_{\tilde{N}_{\{1,2\} i\left(s_{m}, t\right)}(t-1)}\right) \\
& =\sum_{\mathrm{s}_{m} \in \mathrm{~S}_{m}} \prod_{t=1}^{n}\left(\frac{\tilde{N}_{i\left(\mathrm{~s}_{m}, t\right)}(t-1)+1}{t+1}\right)  \tag{3}\\
& =\frac{1}{(n+1)!} \sum_{\mathrm{s}_{m} \in \mathrm{~S}_{m}} \prod_{t=1}^{n}\left(\tilde{N}_{i\left(\mathrm{~s}_{m}, t\right)}(t-1)+1\right) \\
& =\frac{1}{(n+1)!} \sum_{\mathrm{s}_{m} \in \mathrm{~S}_{m}} m!(n-m)!,
\end{align*}
$$

where the last equality follows since $\tilde{N}_{1}(n)=m, \tilde{N}_{2}(n)=n-m, \tilde{N}_{1}(0)=\tilde{N}_{2}(0)=0$ on $\mathrm{s}_{m}$. Therefore, we have for all $m \in\{0, \ldots, n\}$ that

$$
\begin{equation*}
\mathbb{P}\left(N_{1}(n)=m\right)=\frac{\binom{n}{m} m!(n-m)!}{(n+1)!}=\frac{n!}{(n+1)!}=\frac{1}{n+1} . \tag{A.30}
\end{equation*}
$$

This, in fact, proves a stronger result that $N_{1}(n) / n$ is uniformly distributed on $\left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1\right\}$ for any $n \in \mathbb{N}$. The desired result now follows as a corollary in the limit $n \rightarrow \infty$; for an arbitrary $x \in[0,1]$, consider

$$
\mathbb{P}\left(\frac{N_{1}(n)}{n} \leqslant x\right)=\sum_{m=0}^{\lfloor x n\rfloor} \mathbb{P}\left(N_{1}(n)=m\right)=\frac{\lfloor x n\rfloor+1}{n+1},
$$

where the last equality follows using (A.30). Thus, we have $\lim _{n \rightarrow \infty} \mathbb{P}\left(\frac{N_{1}(n)}{n} \leqslant x\right)=x$ for any $x \in[0,1]$, i.e., $\frac{N_{1}(n)}{n}$ converges in law to the Uniform distribution on $[0,1]$.

## A. 7 Proof of Theorem 4

We essentially need to bound the growth rate of $R_{n}^{\pi}$ under the policy $\pi$ given by Algorithm 1 with $\rho>1$, in three (exhaustive) regimes, viz., (i) $\Delta=o\left(\sqrt{\frac{\log n}{n}}\right)$ ("small gap"), (ii) $\Delta=\omega\left(\sqrt{\frac{\log n}{n}}\right)$ ("large gap"), and (iii) $\Delta=\Theta\left(\sqrt{\frac{\log n}{n}}\right)$ ("moderate gap"). We handle the three cases below separately.

## A.7.1 The "small gap" regime

Here, we have

$$
\frac{\mathbb{E} R_{n}^{\pi}}{\sqrt{n \log n}} \leqslant \frac{\Delta n}{\sqrt{n \log n}}=\sqrt{\frac{\Delta^{2} n}{\log n}} .
$$

Since $\Delta=o\left(\sqrt{\frac{\log n}{n}}\right)$, it follows that $\mathbb{E} R_{n}^{\pi}=o(\sqrt{n \log n})$. Therefore, we conclude using Markov's inequality that $R_{n}^{\pi}=o_{p}(\sqrt{n \log n})$ whenever $\Delta=o\left(\sqrt{\frac{\log n}{n}}\right)$.

## A.7.2 The "large gap" regime

In this regime, we have

$$
\frac{\mathbb{E} R_{n}^{\pi}}{\sqrt{n \log n}} \leqslant \frac{C \rho\left(\frac{\log n}{\Delta}+\frac{\Delta}{\rho-1}\right)}{\sqrt{n \log n}},
$$

where $C$ is some absolute constant (follows from [4], Theorem 7). Since $\Delta=\omega\left(\sqrt{\frac{\log n}{n}}\right)$ and $\Delta \leqslant 1$ (rewards bounded in $[0,1])$, it follows that $\mathbb{E} R_{n}^{\pi}=o(\sqrt{n \log n})$. Thus, we again conclude using Markov's inequality that $R_{n}^{\pi}=o_{p}(\sqrt{n \log n})$ whenever $\Delta=\omega\left(\sqrt{\frac{\log n}{n}}\right)$.

## A.7.3 The "moderate gap" regime

Since $\Delta=\Theta\left(\sqrt{\frac{\log n}{n}}\right)$, there exists some $\theta \in \mathbb{R}_{+}$and a diverging sequence of natural numbers $\left\{n_{k}\right\}_{k \in \mathbb{N}}$ such that $\Delta$ scales with the horizon of play $n_{k}$ along this sequence as $\Delta=\sqrt{\frac{\theta \log n_{k}}{n_{k}}}$. Without loss of generality, suppose that arm 1 is optimal, i.e., $\mu_{1} \geqslant \mu_{2}$. We then have

$$
\begin{align*}
\frac{R_{n_{k}}^{\pi}}{\sqrt{n_{k} \log n_{k}}} & =\frac{\mu_{1} n_{k}-\sum_{j=1}^{N_{1}\left(n_{k}\right)} X_{1, j}-\sum_{j=1}^{N_{2}\left(n_{k}\right)} X_{2, j}}{\sqrt{n_{k} \log n_{k}}}  \tag{1.1}\\
& =\frac{\mu_{1} n_{k}-\mu_{1} N_{1}\left(n_{k}\right)-\mu_{2} N_{2}\left(n_{k}\right)-\sum_{j=1}^{N_{1}\left(n_{k}\right)} Y_{1, j}-\sum_{j=1}^{N_{2}\left(n_{k}\right)} Y_{2, j}}{\sqrt{n_{k} \log n_{k}}},
\end{align*}
$$

where $Y_{i, j}:=X_{i, j}-\mu_{i}, i \in\{1,2\}, j \in \mathbb{N}$. Therefore,

$$
\begin{align*}
\frac{R_{n_{k}}^{\pi}}{\sqrt{n_{k} \log n_{k}}} & =\frac{\Delta N_{2}\left(n_{k}\right)-\sum_{j=1}^{N_{1}\left(n_{k}\right)} Y_{1, j}-\sum_{j=1}^{N_{2}\left(n_{k}\right)} Y_{2, j}}{\sqrt{n_{k} \log n_{k}}} \\
& =\Delta \sqrt{\frac{n_{k}}{\log n_{k}}}\left(\frac{N_{2}\left(n_{k}\right)}{n_{k}}\right)-\left(\frac{\sum_{j=1}^{N_{1}\left(n_{k}\right)} Y_{1, j}+\sum_{j=1}^{N_{2}\left(n_{k}\right)} Y_{2, j}}{\sqrt{n_{k} \log n_{k}}}\right) \\
& =\sqrt{\theta}\left(\frac{N_{2}\left(n_{k}\right)}{n_{k}}\right)-\left(\frac{\sum_{j=1}^{N_{1}\left(n_{k}\right)} Y_{1, j}+\sum_{j=1}^{N_{2}\left(n_{k}\right)} Y_{2, j}}{\sqrt{n_{k} \log n_{k}}}\right) . \tag{A.31}
\end{align*}
$$

Consider the summation terms above. We have

$$
\begin{align*}
\left|\frac{\sum_{j=1}^{N_{1}\left(n_{k}\right)} Y_{1, j}+\sum_{j=1}^{N_{2}\left(n_{k}\right)} Y_{2, j}}{\sqrt{n_{k} \log n_{k}}}\right| & \leqslant\left|\frac{\sum_{j=1}^{N_{1}\left(n_{k}\right)} Y_{1, j}}{\sqrt{n_{k} \log n_{k}}}\right|+\left|\frac{\sum_{j=1}^{N_{2}\left(n_{k}\right)} Y_{2, j}}{\sqrt{n_{k} \log n_{k}}}\right| \\
& \leqslant\left|\frac{\sum_{j=1}^{N_{1}\left(n_{k}\right)} Y_{1, j}}{\sqrt{N_{1}\left(n_{k}\right) \log N_{1}\left(n_{k}\right)}}\right|+\left|\frac{\sum_{j=1}^{N_{2}\left(n_{k}\right)} Y_{2, j}}{\sqrt{N_{2}\left(n_{k}\right) \log N_{2}\left(n_{k}\right)}}\right| . \tag{A.32}
\end{align*}
$$

Since $N_{i}\left(n_{k}\right)$, for each $i \in\{1,2\}$, can be lower-bounded path-wise by a deterministic monotone increasing coercive function of $k$, say $f(k)$, we have

$$
\begin{equation*}
\left|\frac{\sum_{j=1}^{N_{i}\left(n_{k}\right)} Y_{i, j}}{\sqrt{N_{i}\left(n_{k}\right) \log N_{i}\left(n_{k}\right)}}\right| \leqslant \sup _{m \geqslant f(k)}\left|\frac{\sum_{j=1}^{m} Y_{i, j}}{\sqrt{m \log m}}\right| \quad \forall i \in\{1,2\} . \tag{A.33}
\end{equation*}
$$

Since $Y_{i, j}$ 's are independent, zero-mean, bounded random variables, it follows from the Law of the Iterated Logarithm (see [25], Theorem 8.5.2) that

$$
\begin{equation*}
\sup _{m \geqslant f(k)}\left|\frac{\sum_{j=1}^{m} Y_{i, j}}{\sqrt{m \log m}}\right| \xrightarrow[k \rightarrow \infty]{\text { w.p. } 1} 0 \quad \forall i \in\{1,2\} . \tag{A.34}
\end{equation*}
$$

Combining (A.31), (A.32), (A.33) and (A.34), we conclude

$$
\frac{R_{n_{k}}^{\pi}}{\sqrt{n_{k} \log n_{k}}}=\sqrt{\theta}\left(\frac{N_{2}\left(n_{k}\right)}{n_{k}}\right)+o_{p}(1)
$$

From Theorem 1, we know that when $\Delta \sim \sqrt{\frac{\theta \log n_{k}}{n_{k}}}, \frac{N_{2}\left(n_{k}\right)}{n_{k}} \xrightarrow[k \rightarrow \infty]{p} 1-\lambda_{\rho}^{*}(\theta)$. Thus, it follows that

$$
\frac{R_{n_{k}}^{\pi}}{\sqrt{n_{k} \log n_{k}}} \stackrel{p}{k \rightarrow \infty} \sqrt{\theta}\left(1-\lambda_{\rho}^{*}(\theta)\right)=h_{\rho}(\theta) .
$$

Since $\theta \in \mathbb{R}_{+}$is arbitrary, the worst-case regret in the $\Delta=\Theta\left(\sqrt{\frac{\log n}{n}}\right)$ regime corresponds to the choice of $\theta$ given by $\theta_{\rho}^{*}=\arg \max _{\theta \geqslant 0} h_{\rho}(\theta)$. Since we already know that $R_{n}^{\pi}=o_{p}(\sqrt{n \log n})$ in the other two regimes ("small" and "large" gaps), it must be that the $\theta_{\rho}^{*}$ so obtained indeed corresponds to the global (in $\Delta$ ) worst-case regret of Algorithm 1.

## A. 8 Proof of Theorem 5

Notation. Let $C$ be the space of continuous functions $[0,1] \mapsto \mathbb{R}^{2}$, endowed with the uniform metric. Let $\mathcal{D}$ be the space of right-continuous functions with left limits, mapping $[0,1] \mapsto \mathbb{R}^{2}$, and endowed with the Skorohod metric (see [70], Chapters 2 and 3, for an overview). Let $\mathcal{D}_{0}$ be
the set of elements of $\mathcal{D}$ of the form $\left(\phi_{1}, \phi_{2}\right)$, where $\phi_{i}$ is a non-decreasing real-valued function satisfying $0 \leqslant \phi_{i}(t) \leqslant 1$ for $i \in\{1,2\}$ and $t \in[0,1]$. For $t \in[0,1]$, denote the identity map by $\mathfrak{e}(t):=t$.

For $i \in\{1,2\}$ and $t \in[0,1]$, define $\Psi_{i, n}(t):=\frac{\sum_{j=1}^{\lfloor n t\rfloor} X_{i, j}-\mu n t}{\sqrt{n}}$. Then, $\left(\Psi_{1, n}, \Psi_{2, n}\right) \in \mathcal{D}$. Also for $i \in\{1,2\}$ and $t \in[0,1]$, define $W_{i}(t):=\theta_{i} t+\sigma_{i} B_{i}^{\prime}(t)$, where $B_{1}^{\prime}$ and $B_{2}^{\prime}$ are independent standard Brownian motions in $\mathbb{R}$. Note that $\mathbb{P}\left(W_{i} \in C\right)=1$ for $i \in\{1,2\}$. Since $\left(X_{i, j}\right)_{i \in\{1,2\}, j \in \mathbb{N}}$ 's are independent random variables (i.i.d. within and independent across sequences), we know from Donsker's Theorem (see [70], Section 14, for details) that as $n \rightarrow \infty$,

$$
\left(\Psi_{1, n}, \Psi_{2, n}\right) \Rightarrow\left(W_{1}, W_{2}\right) \text { in } \mathcal{D} .
$$

For $i \in\{1,2\}$ and $t \in[0,1]$, define $\phi_{i, n}(t):=\frac{N_{i}(\lfloor n t\rfloor)}{n}$. Thus, $\left(\phi_{1, n}, \phi_{2, n}\right) \in \mathcal{D}_{0}$, and it follows from the result for the "small gap" regime in Theorem 1 that as $n \rightarrow \infty$,

$$
\left(\phi_{1, n}, \phi_{2, n}\right) \xrightarrow{p}\left(\frac{\mathfrak{e}}{2}, \frac{\mathfrak{e}}{2}\right) \text { in } \mathcal{D}_{0} .
$$

Thus, we have convergence in the product space (see [70], Theorem 3.9), i.e., as $n \rightarrow \infty$,

$$
\left(\Psi_{1, n}, \Psi_{2, n}, \phi_{1, n}, \phi_{2, n}\right) \Rightarrow\left(W_{1}, W_{2}, \frac{\mathfrak{e}}{2}, \frac{\mathfrak{e}}{2}\right) \text { in } \mathcal{D} \times \mathcal{D}_{0}
$$

For $i \in\{1,2\}$ and $t \in[0,1]$, define the composition $\left(\Psi_{i, n} \circ \phi_{i, n}\right)(t):=\Psi_{i, n}\left(\phi_{i, n}(t)\right)$, and $\left(W_{i} \circ \frac{\mathfrak{e}}{2}\right)(t):=W_{i}\left(\frac{\mathfrak{e}(t)}{2}\right)=W_{i}\left(\frac{t}{2}\right)$. Since $W_{1}, W_{2}, \mathfrak{e} \in C$ w.p. 1, it follows from the random timechange lemma (see [70], Section 14, for details) that as $n \rightarrow \infty$

$$
\left(\Psi_{1, n} \circ \phi_{1, n}, \Psi_{2, n} \circ \phi_{2, n}\right) \Rightarrow\left(W_{1} \circ \frac{\mathfrak{e}}{2}, W_{2} \circ \frac{\mathfrak{e}}{2}\right) \text { in } \mathcal{D} .
$$

The stated assertion on cumulative rewards now follows by recognizing for $i \in\{1,2\}$ and
$t \in[0,1]$ that $\left(\Psi_{i, n} \circ \phi_{i, n}\right)(t)=\frac{\tilde{S}_{i,[n t]}}{\sqrt{n}}$, and defining $B_{i}(t):=\sqrt{2} B_{i}^{\prime}\left(\frac{t}{2}\right)$. To prove the assertion on regret, assume without loss of generality that arm 1 is optimal, i.e., $\theta_{1} \geqslant \theta_{2}$. Then, the result follows after a direct application of the Continuous Mapping Theorem (see [70], Theorem 2.7), to wit,

$$
R_{\lfloor n t\rfloor}^{\pi}=\left(\mu+\frac{\theta_{1}}{\sqrt{n}}\right)\lfloor n t\rfloor-S_{1,\lfloor n t\rfloor}-S_{2,\lfloor n t\rfloor}=\frac{\theta_{1}\lfloor n t\rfloor}{\sqrt{n}}-\left(\tilde{S}_{1,\lfloor n t\rfloor}+\tilde{S}_{2,\lfloor n t\rfloor}\right),
$$

and therefore as $n \rightarrow \infty$,
$\left(\frac{R_{\lfloor n t\rfloor}^{\pi}}{\sqrt{n}}\right)_{t \in[0,1]} \Rightarrow\left(\theta_{1} t-\left(\left(\frac{\theta_{1}+\theta_{2}}{2}\right) t+\frac{\sigma_{1}}{\sqrt{2}} B_{1}(t)+\frac{\sigma_{2}}{\sqrt{2}} B_{2}(t)\right)\right)_{t \in[0,1]}=\left(\frac{\Delta_{0} t}{2}+\sqrt{\frac{\sigma_{1}^{2}+\sigma_{2}^{2}}{2}} \tilde{B}(t)\right)_{t \in[0,1]}$,
where $\tilde{B}(t):=-\sqrt{\frac{\sigma_{1}^{2}}{\sigma_{1}^{2}+\sigma_{2}^{2}}} B_{1}(t)-\sqrt{\frac{\sigma_{2}^{2}}{\sigma_{1}^{2}+\sigma_{2}^{2}}} B_{2}(t)$.

## A. 9 Proof of Theorem 2

We will prove this result in two parts; the preamble in A.9.1 below will prove a meta-result stating that $N_{i}(n) / n>1 /(2|\mathcal{I}|)$ with high probability (approaching 1 as $n \rightarrow \infty$ ) for any arm $i \in$ $\mathcal{I}$. We will then leverage this meta-result to prove the assertions of the theorem in A.9.2.

## A.9.1 Preamble

Let $L:=|I|$. If $L=1$, the result follows trivially from the standard logarithmic bound for the expected regret (Theorem 7 in [4]), followed by Markov's inequality. Therefore, without loss of generality, suppose that $|\mathcal{I}| \geqslant 2$, and fix an arbitrary arm $i \in \mathcal{I}$. Then, we know that the following
is true for any integer $u>1$ :

$$
\begin{aligned}
N_{i}(n) & \leqslant u+\sum_{t=u}^{n-1} \mathbb{1}\left\{\pi_{t+1}=i, N_{i}(t) \geqslant u\right\} \\
& \leqslant u+\sum_{t=u}^{n-1} \mathbb{1}\left\{\pi_{t+1}=i, N_{i}(t) \geqslant u, \sum_{j \in \mathcal{T} \backslash\{i\}} N_{j}(t) \leqslant t-u\right\},
\end{aligned}
$$

where $\pi_{t+1} \in[K]$ indicates the arm played at time $t+1$. In particular, the above holds also for $u=\left\lceil\left(\frac{1}{L}+\epsilon\right) n\right\rceil$, where $\epsilon \in\left(0, \frac{L-1}{L}\right)$ is arbitrarily chosen. We will fix this $u$ going forward, even though we may not always express its value explicitly for readability of the analysis that follows. We thus have

$$
\begin{align*}
N_{i}(n) & \leqslant u+\sum_{t=u}^{n-1} \mathbb{1}\left\{B_{i, N_{i}(t), t} \geqslant \max _{\hat{j} \in[K] \backslash i\}} B_{\hat{j}, N_{\hat{j}}(t), t}, N_{i}(t) \geqslant u, \sum_{j \in \mathcal{I} \backslash\{i\}} N_{j}(t) \leqslant t-u\right\} \\
& \leqslant u+\sum_{t=u}^{n-1} \mathbb{1}\left\{B_{i, N_{i}(t), t} \geqslant \max _{\hat{j} \in \mathcal{I} \backslash\{i\}} B_{\hat{j}, N_{\hat{j}}(t), t}, N_{i}(t) \geqslant u, \sum_{j \in \mathcal{I} \backslash\{i\}} N_{j}(t) \leqslant t-u\right\}, \tag{A.35}
\end{align*}
$$

where $B_{k, s, t}:=\hat{X}_{k}(s)+\sqrt{(\rho \log t) / s}$ for $k \in[K]$, and $\hat{X}_{k}(s):=\sum_{l=1}^{s} X_{k, l} / s$ denotes the empirical mean reward from the "first splays" of arm $k$ (Note the distinction from $\bar{X}_{k}(s)$, which has been defined before as the empirical mean reward of arm $k$ "at time $s$," i.e., mean over its "first $N_{k}(s)$ plays"). Now observe that

$$
\begin{equation*}
\left\{\sum_{j \in \mathcal{I} \backslash\{i\}} N_{j}(t) \leqslant t-u\right\} \subseteq\left\{\exists j \in \mathcal{I} \backslash\{i\}: N_{j}(t) \leqslant \frac{t-u}{L-1}\right\} \subseteq\left\{\exists j \in \mathcal{I} \backslash\{i\}: N_{j}(t) \leqslant\left(\frac{1}{L}-\frac{\epsilon}{L-1}\right) t\right\}, \tag{A.36}
\end{equation*}
$$

where the last inclusion follows using $u=\left\lceil\left(\frac{1}{L}+\epsilon\right) n\right\rceil$ and $n \geqslant t$. Combining (A.35) and (A.36) using the Union bound, we obtain

$$
\begin{align*}
N_{i}(n) & \leqslant u+\sum_{t=u}^{n-1} \sum_{j \in \mathcal{T} \backslash\{i\}} \mathbb{1}\left\{B_{i, N_{i}(t), t} \geqslant \max _{\hat{j} \in I \backslash\{i\}} B_{\hat{j}, N_{\hat{j}}(t), t}, N_{i}(t) \geqslant u, N_{j}(t) \leqslant\left(\frac{1}{L}-\frac{\epsilon}{L-1}\right) t\right\} \\
& \leqslant u+\sum_{t=u}^{n-1} \sum_{j \in \mathcal{T} \backslash\{i\}} \mathbb{1}\left\{B_{i, N_{i}(t), t} \geqslant B_{j, N_{j}(t), t}, N_{i}(t) \geqslant u, N_{j}(t) \leqslant\left(\frac{1}{L}-\frac{\epsilon}{L-1}\right) t\right\} \\
& \leqslant u+\underbrace{\sum_{t=u}^{n-1} \sum_{j \in \mathcal{T} \backslash\{i\}} \mathbb{1}\left\{B_{i, N_{i}(t), t} \geqslant B_{j, N_{j}(t), t}, N_{i}(t) \geqslant\left(\frac{1}{L}+\epsilon\right) t, N_{j}(t) \leqslant\left(\frac{1}{L}-\frac{\epsilon}{L-1}\right) t\right\}}_{=: Z_{n}}, \tag{A.37}
\end{align*}
$$

where the last inequality again uses $u=\left\lceil\left(\frac{1}{L}+\epsilon\right) n\right\rceil$ and $n \geqslant t$. Define the events:
$E_{i}:=\left\{N_{i}(t) \geqslant\left(\frac{1}{L}+\epsilon\right) t\right\}$, and $E_{j}:=\left\{N_{j}(t) \leqslant\left(\frac{1}{L}-\frac{\epsilon}{L-1}\right) t\right\}$. Now,

$$
\begin{align*}
& \mathbb{E} Z_{n} \\
= & \sum_{t=u}^{n-1} \sum_{j \in \mathcal{I} \backslash\{i\}} \mathbb{P}\left(B_{i, N_{i}(t), t} \geqslant B_{j, N_{j}(t), t}, E_{i}, E_{j}\right) \\
= & \sum_{t=u}^{n-1} \sum_{j \in \mathcal{T} \backslash i\}} \mathbb{P}\left(\hat{Y}_{i}\left(N_{i}(t)\right)-\hat{Y}_{j}\left(N_{j}(t)\right) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{j}(t)}}-\frac{1}{\sqrt{N_{i}(t)}}\right), E_{i}, E_{j}\right), \tag{A.38}
\end{align*}
$$

where $\hat{Y}_{k}(s):=\sum_{l=1}^{s} Y_{k, l} / s$ and $Y_{k, l}:=X_{k, l}-\mathbb{E} X_{k, l}$ for $k \in[K], s \in \mathbb{N}, l \in \mathbb{N}$. The last equality above follows since $i, j \in I$ and the mean rewards of arms in $I$ are equal. Thus,

$$
\mathbb{E} Z_{n} \leqslant \sum_{t=u}^{n-1} \sum_{j \in \mathcal{I} \backslash\{i\}} \sum_{m_{i}=\left\lceil\left(\frac{1}{L}+\epsilon\right) t\right\rceil}^{t} \sum_{m_{j}=1}^{\left\lfloor\left(\frac{1}{L}-\frac{\epsilon}{L-1}\right) t\right\rfloor} \mathbb{P}\left(\hat{Y}_{i}\left(m_{i}\right)-\hat{Y}_{j}\left(m_{j}\right) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{m_{j}}}-\frac{1}{\sqrt{m_{i}}}\right)\right) .
$$

Since $\mathbb{E}\left[\hat{Y}_{i}\left(m_{i}\right)-\hat{Y}_{j}\left(m_{j}\right)\right]=0$, and $m_{j}<m_{i}$ over the range of the summation above, we can use the Chernoff-Hoeffding bound (Fact 1) to obtain

$$
\begin{equation*}
\mathbb{E} Z_{n} \leqslant \sum_{t=u}^{n-1} \sum_{j \in \mathcal{I} \backslash\{i\}} \sum_{m_{i}=\left\lceil\left(\frac{1}{L}+\epsilon\right) t\right\rceil}^{t} \sum_{m_{j}=1}^{\left\lfloor\left(\frac{1}{L}-\frac{\epsilon}{L-1}\right) t\right\rfloor} \exp \left[-2 \rho\left(1-2 \sqrt{\frac{m_{i} m_{j}}{\left(m_{i}+m_{j}\right)^{2}}}\right) \log t\right] . \tag{A.39}
\end{equation*}
$$

Let $\gamma:=m_{i} /\left(m_{i}+m_{j}\right)$. Then, $\gamma \geqslant \frac{\frac{1}{L}+\epsilon}{\frac{L}{L}+\left(\frac{L-2}{L-1}\right) \epsilon}>1 / 2$ over the range of the summation in (A.39). Consequently, $\gamma(1-\gamma)$ is maximized at $\gamma=\frac{\frac{1}{L}+\epsilon}{\frac{2}{L}+\left(\frac{L-2}{L-1}\right) \epsilon}$, and therefore, $m_{i} m_{j} /\left(m_{i}+m_{j}\right)^{2}=\gamma(1-$ $\gamma) \leqslant(f(\epsilon, L))^{2}<1 / 4$ in (A.39), where $f(\epsilon, L)$ as defined as:

$$
\begin{equation*}
f(\epsilon, L):=\sqrt{\frac{(L-1)(1+\epsilon L)(L-1-\epsilon L)}{(2(L-1)+L(L-2) \epsilon)^{2}}} . \tag{A.40}
\end{equation*}
$$

Combining (A.39) and (A.40), we obtain

$$
\begin{equation*}
\mathbb{E} Z_{n} \leqslant(L-1) \sum_{t=u}^{n-1} t^{-2(\rho-1-2 \rho f(\epsilon, L))} \tag{A.41}
\end{equation*}
$$

Now consider an arbitrary $\delta>0$. From (A.37), we have

$$
\begin{equation*}
\mathbb{P}\left(N_{i}(n) \geqslant u+\delta n\right) \leqslant \mathbb{P}\left(Z_{n} \geqslant \delta n\right) \underset{(\star)}{\leqslant} \frac{\mathbb{E} Z_{n}}{\delta n} \underset{(\dagger)}{\leqslant}\left(\frac{L-1}{\delta n}\right) \sum_{t=u}^{n-1} t^{-2(\rho(1-2 f(\epsilon, L))-1)}, \tag{A.42}
\end{equation*}
$$

where $(\star)$ is due to Markov's inequality, and $(\dagger)$ follows using (A.41). Observe from (A.40) that $f(\epsilon, L)$ is monotone decreasing in $\epsilon$ over the interval $\epsilon \in\left[0, \frac{L-1}{L}\right]$, with $f(0, L)=1 / 2$ and $f\left(\frac{L-1}{L}, L\right)=0$. Therefore, $1-2 f(\epsilon, L)>0$ in the interval $\epsilon \in\left(0, \frac{L-1}{L}\right]$. Thus, for $\rho$ large enough, the exponent of $t$ in (A.42) can be made arbitrarily small. That is, $\exists \rho_{0} \in \mathbb{R}_{+}$s.t. for all $\rho \geqslant \rho_{0}$, we have $2(\rho(1-2 f(\epsilon, L))-1)>0 \forall \epsilon \in\left[\frac{1}{2 L(L-1)}, \frac{L-1}{L}\right]$. Now supposing $\rho \geqslant \rho_{0}$, plug in $\epsilon=\frac{1}{2 L(L-1)}$ in (A.42) (this includes substituting $u=\left\lceil\left(\frac{1}{L}+\frac{1}{2 L(L-1)}\right) n\right\rceil$ ). Then since $\delta>0$ and
$i \in \mathcal{I}$ are arbitrary, it follows that for any $\delta>0$ and $i \in \mathcal{I}$,

$$
\begin{equation*}
\lim _{n \rightarrow \infty} \mathbb{P}\left(N_{i}(n) \geqslant\left(\frac{1}{L}+\frac{1}{2 L(L-1)}+\delta\right) n\right)=\frac{L^{2 \rho}}{\delta} \lim _{n \rightarrow \infty} n^{-2\left(\rho\left(1-2 f\left(\frac{1}{2 L(L-1)}, L\right)\right)-1\right)}=0 \tag{A.43}
\end{equation*}
$$

Notice that for any $\delta>0$ and $i \in \mathcal{I}$,

$$
\begin{aligned}
& \mathbb{P}\left(N_{i}(n)+\sum_{j \in[K] \backslash I} N_{j}(n) \leqslant\left(\frac{1}{2 L}-(L-1) \delta\right) n\right) \\
= & \mathbb{P}\left(\sum_{j \in \mathcal{I} \backslash\{i\}} N_{j}(n) \geqslant(L-1)\left(\frac{1}{L}+\frac{1}{2 L(L-1)}+\delta\right) n\right) \\
\leqslant & \sum_{j \in \mathcal{I} \backslash\{i\}} \mathbb{P}\left(N_{j}(n) \geqslant\left(\frac{1}{L}+\frac{1}{2 L(L-1)}+\delta\right) n\right),
\end{aligned}
$$

where the last inequality follows using the Union bound. Taking limits on both sides above, we conclude using (A.43) that for any $\delta>0$ and $i \in I$,

$$
\begin{equation*}
\lim _{n \rightarrow \infty} \mathbb{P}\left(N_{i}(n)+\sum_{j \in[K] \backslash I} N_{j}(n) \leqslant\left(\frac{1}{2 L}-(L-1) \delta\right) n\right)=0 . \tag{A.44}
\end{equation*}
$$

If $\mathcal{I}=[K]$, the conclusion that $N_{i}(n) / n>1 /(2 L)=1 /(2 K)$ with high probability (approaching 1 as $n \rightarrow \infty$ ) for all $i \in[K]$, is immediate from (A.44). If $\mathcal{I} \neq[K]$, then $\sum_{j \in[K] \backslash I} \mathbb{E}\left(N_{j}(n) / n\right) \leqslant$ $C K \rho\left[\left(\frac{1}{\Delta_{\min }^{2}}\right)\left(\frac{\log n}{n}\right)+\frac{1}{(\rho-1) n}\right]$ for some absolute constant $C>0$ follows from [4], Theorem 7. Consequently if $\Delta_{\min }=\omega\left(\sqrt{\frac{\log n}{n}}\right)$, Markov's inequality implies that $\sum_{j \in[K] \backslash I} N_{j}(n) / n=o_{p}(1)$. Thus, it again follows using (A.44) that $N_{i}(n) / n>1 /(2 L)$ with high probability (approaching 1 as $n \rightarrow \infty$ ) for all $i \in I$.

## A.9.2 Proof of part (I) and (II)

Note that the following holds for any integer $u>1$ and any arm $i \in[K]$ :

$$
N_{i}(n) \leqslant u+\sum_{t=u+1}^{n} \mathbb{1}\left\{\pi_{t}=i, N_{i}(t-1) \geqslant u\right\},
$$

where $\pi_{t} \in[K]$ indicates the arm played at time $t$. In particular, the above is true also for $u=$ $N_{j}(n)+\lceil\epsilon n\rceil$, where $j \in[K] \backslash\{i\}$ and $\epsilon>0$ are arbitrarily chosen. Without loss of generality, suppose that $|\mathcal{I}| \geqslant 2$ (the result is trivial for $|\mathcal{I}|=1$ ), and fix two arbitrary arms $i, j \in \mathcal{I}$. Then,

$$
\begin{aligned}
N_{i}(n) & \leqslant N_{j}(n)+\lceil\epsilon n\rceil+\sum_{t=N_{j}(n)+\lceil\epsilon n\rceil+1}^{n} \mathbb{1}\left\{\pi_{t}=i, N_{i}(t-1) \geqslant N_{j}(n)+\lceil\epsilon n\rceil\right\} \\
& \leqslant N_{j}(n)+\lceil\epsilon n\rceil+\sum_{t=\lceil\epsilon n\rceil+1}^{n} \mathbb{1}\left\{\pi_{t}=i, N_{i}(t-1) \geqslant N_{j}(n)+\epsilon n\right\} \\
& \leqslant N_{j}(n)+\lceil\epsilon n\rceil+\sum_{t=\lceil\epsilon n\rceil+1}^{n} \mathbb{1}\left\{B_{i, N_{i}(t-1), t-1} \geqslant B_{j, N_{j}(t-1), t-1}, N_{i}(t-1) \geqslant N_{j}(n)+\epsilon n\right\},
\end{aligned}
$$

where $B_{k, s, t}:=\hat{X}_{k}(s)+\sqrt{(\rho \log t) / s}$ for $k \in[K]$, with $\hat{X}_{k}(s)$ denoting the empirical mean reward from the first $s$ plays of arm $k$. Then,

$$
\begin{align*}
N_{i}(n) & \leqslant N_{j}(n)+\lceil\epsilon n\rceil+\sum_{t=\lceil\epsilon n\rceil}^{n-1} \mathbb{1}\left\{B_{i, N_{i}(t), t} \geqslant B_{j, N_{j}(t), t}, N_{i}(t) \geqslant N_{j}(n)+\epsilon n\right\} \\
& \leqslant N_{j}(n)+\lceil\epsilon n\rceil+\sum_{t=\lceil\epsilon n\rceil}^{n-1} \mathbb{1}\left\{B_{i, N_{i}(t), t} \geqslant B_{j, N_{j}(t), t}, N_{i}(t) \geqslant N_{j}(t)+\epsilon t\right\} \\
& \leqslant N_{j}(n)+\epsilon n+1+Z_{n}, \tag{A.45}
\end{align*}
$$

where $Z_{n}:=\sum_{t=[\epsilon n]}^{n-1} \mathbb{1}\left\{B_{i, N_{i}(t), t} \geqslant B_{j, N_{j}(t), t}, N_{i}(t) \geqslant N_{j}(t)+\epsilon t\right\}$. Now,

$$
\begin{aligned}
& \mathbb{E} Z_{n} \\
= & \sum_{t=\lceil\epsilon n\rceil}^{n-1} \mathbb{P}\left\{B_{i, N_{i}(t), t} \geqslant B_{j, N_{j}(t), t}, N_{i}(t) \geqslant N_{j}(t)+\epsilon t\right\} \\
= & \sum_{t=\lceil\epsilon n\rceil}^{n-1} \mathbb{P}\left\{\hat{X}_{i}\left(N_{i}(t)\right)-\hat{X}_{j}\left(N_{j}(t)\right) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{j}(t)}}-\frac{1}{\sqrt{N_{i}(t)}}\right), N_{i}(t) \geqslant N_{j}(t)+\epsilon t\right\} \\
= & \sum_{t=\lceil\epsilon n\rceil}^{n-1} \mathbb{P}\left\{\hat{Y}_{i}\left(N_{i}(t)\right)-\hat{Y}_{j}\left(N_{j}(t)\right) \geqslant \sqrt{\rho \log t}\left(\frac{1}{\sqrt{N_{j}(t)}}-\frac{1}{\sqrt{N_{i}(t)}}\right), N_{i}(t) \geqslant N_{j}(t)+\epsilon t\right\},
\end{aligned}
$$

where $\hat{Y}_{k}(s):=\sum_{l=1}^{s} Y_{k, l} / s$ for $k \in[K], s \in \mathbb{N}$, with $Y_{k, l}:=X_{k, l}-\mathbb{E} X_{k, l}$ for $l \in \mathbb{N}$. The last equality above follows since $i, j \in I$ and the mean rewards of arms in $I$ are equal. Thus,

$$
\begin{align*}
& \mathbb{E} Z_{n} \\
= & \sum_{t=\lceil\epsilon n\rceil}^{n-1} \mathbb{P}\left(\hat{Y}_{i}\left(N_{i}(t)\right)-\hat{Y}_{j}\left(N_{j}(t)\right) \geqslant \sqrt{\frac{\rho \log t}{N_{i}(t)}}\left(\sqrt{\frac{N_{i}(t)}{N_{j}(t)}}-1\right), N_{i}(t) \geqslant N_{j}(t)+\epsilon t\right) \\
\leqslant & \sum_{t=\lceil\epsilon n\rceil}^{n-1} \mathbb{P}\left(\hat{Y}_{i}\left(N_{i}(t)\right)-\hat{Y}_{j}\left(N_{j}(t)\right) \geqslant \sqrt{\frac{\rho \log t}{t}}(\sqrt{1+\epsilon}-1), N_{i}(t) \geqslant N_{j}(t)+\epsilon t\right) \\
\leqslant & \sum_{t=\lceil\epsilon n\rceil}^{n-1} \mathbb{P}\left(W_{t} \geqslant \sqrt{1+\epsilon}-1\right), \tag{A.46}
\end{align*}
$$

where $W_{t}:=\sqrt{\frac{t}{\rho \log t}}\left(\frac{\sum_{l=1}^{N_{i}(t)} Y_{i, l}}{N_{i}(t)}-\frac{\sum_{l=1}^{N_{j}(t)} Y_{j, l}}{N_{j}(t)}\right)$. Now,

$$
\left.\left.\begin{array}{rl} 
& \left|W_{t}\right| \\
\leqslant & \sqrt{\frac{t}{\rho \log t}}\left(\left|\frac{\sum_{l=1}^{N_{i}(t)} Y_{i, l}}{N_{i}(t)}\right|+\left|\frac{\sum_{l=1}^{N_{j}(t)} Y_{j, l}}{N_{j}(t)}\right|\right) \\
=\sqrt{\frac{2 t}{\rho \log t}}\left(\sqrt{\frac{\log \log N_{i}(t)}{N_{i}(t)}} \left\lvert\, \frac{\sum_{l=1}^{N_{i}(t)} Y_{i, l}}{\sqrt{2 N_{i}(t) \log \log N_{i}(t)}}\right.\right. & +\sqrt{\frac{\log \log N_{j}(t)}{N_{j}(t)}} \left\lvert\, \frac{\sum_{l=1}^{N_{j}(t)} Y_{j, j}}{\sqrt{2 N_{j}(t) \log \log N_{j}(t)}}\right. \\
\leqslant & \sqrt{\frac{2 t}{\rho \log t}}\left(\left.\sqrt{\frac{\log \log t}{N_{i}(t)}}\left|\frac{\sum_{l=1}^{N_{i}(t)} Y_{i, l}}{\sqrt{2 N_{i}(t) \log \log N_{i}(t)}}\right|+\sqrt{\frac{\log \log t}{N_{j}(t)}} \right\rvert\, \frac{\sum_{l=1}^{N_{j}(t)} Y_{j, l}}{\sqrt{2 N_{j}(t) \log \log N_{j}(t)}}\right.
\end{array}\right) .\right) .
$$

We know that $N_{k}(t)$, for any arm $k \in[K]$, can be lower-bounded path-wise by a deterministic monotone increasing coercive function of $t$, say $h(t)$. This follows as a trivial consequence of the structure of the policy, and the fact that the rewards are uniformly bounded. Therefore, we have for any arm $k \in \mathcal{I}$ that

$$
\begin{align*}
& \left|\frac{\sum_{l=1}^{N_{k}(t)} Y_{k, l}}{\sqrt{2 N_{k}(t) \log \log N_{k}(t)}}\right| \leqslant \sup _{m \geqslant h(t)}\left|\frac{\sum_{l=1}^{m} Y_{k, l}}{\sqrt{2 m \log \log m}}\right| \\
\Longrightarrow & \limsup _{t \rightarrow \infty}\left|\frac{\sum_{l=1}^{N_{k}(t)} Y_{k, l}}{\sqrt{2 N_{k}(t) \log \log N_{k}(t)}}\right| \leqslant \limsup _{t \rightarrow \infty}\left|\frac{\sum_{l=1}^{t} Y_{k, l}}{\sqrt{2 t \log \log t}}\right| . \tag{A.48}
\end{align*}
$$

For any $k \in I$, we know that $\left\{Y_{k, l}: l \in \mathbb{N}\right\}$ is a collection of i.i.d. random variables with $\mathbb{E} Y_{k, 1}=0$ and $\operatorname{Var}\left(Y_{k, 1}\right)=\operatorname{Var}\left(X_{k, 1}\right) \leqslant 1$. Therefore, we conclude using the Law of the Iterated Logarithm (see Theorem 8.5.2 in [25]) in (A.48) that

$$
\begin{equation*}
\limsup _{t \rightarrow \infty}\left|\frac{\sum_{l=1}^{N_{k}(t)} Y_{k, l}}{\sqrt{2 N_{k}(t) \log \log N_{k}(t)}}\right| \leqslant 1 \quad \text { w.p. } 1 \forall k \in \mathcal{I} \text {. } \tag{A.49}
\end{equation*}
$$

Using (A.47), (A.49), and the meta-result from the preamble in A.9.1 that $N_{k}(t) / t>1 /(2|\mathcal{I}|)$
with high probability (approaching 1 as $t \rightarrow \infty$ ) for any arm $k \in \mathcal{I}$, we conclude that

$$
\begin{equation*}
W_{t} \xrightarrow{p} 0 \text { as } t \rightarrow \infty . \tag{A.50}
\end{equation*}
$$

Now,

$$
\mathbb{P}\left(\frac{N_{i}(n)-N_{j}(n)}{n} \geqslant 2 \epsilon\right) \underset{(\dagger)}{\leqslant} \mathbb{P}\left(1+Z_{n} \geqslant \epsilon n\right) \underset{(\ddagger)}{\leqslant} \frac{1+\mathbb{E} Z_{n}}{\epsilon n} \underset{(\star)}{\leqslant} \frac{1}{\epsilon n}+\frac{1}{\epsilon n} \sum_{t=[\epsilon n]}^{n-1} \mathbb{P}\left(W_{t}>\sqrt{1+\epsilon}-1\right),
$$

where $(\dagger)$ follows using (A.45), $(\ddagger)$ using Markov’s inequality, and ( $\star$ ) from (A.46). Therefore,

$$
\begin{equation*}
\mathbb{P}\left(\frac{N_{i}(n)-N_{j}(n)}{n} \geqslant 2 \epsilon\right) \leqslant \frac{1}{\epsilon n}+\left(\frac{1-\epsilon}{\epsilon}\right) \sup _{t \geqslant \epsilon n} \mathbb{P}\left(W_{t}>\sqrt{1+\epsilon}-1\right) . \tag{A.51}
\end{equation*}
$$

Since $\epsilon>0$ is arbitrary, we conclude using (A.50) and (A.51) that for any $\epsilon>0$,

$$
\begin{equation*}
\lim _{n \rightarrow \infty} \mathbb{P}\left(\frac{N_{i}(n)-N_{j}(n)}{n} \geqslant 2 \epsilon\right)=0 \tag{A.52}
\end{equation*}
$$

Our proof is symmetric w.r.t. the labels $i, j$, therefore, an identical result holds also with the labels interchanged in (A.52). Thus, we have $N_{i}(n) / n-N_{j}(n) / n \xrightarrow{p} 0$. Since $i, j$ are arbitrary in $I$, the aforementioned convergence holds for any pair of arms in $I$. Now if $I=[K]$, we are done. If $\mathcal{I} \neq[K]$, then $\sum_{i \in[K] \backslash I} \mathbb{E}\left(N_{i}(n) / n\right) \leqslant C K \rho\left[\left(\frac{1}{\Delta_{\text {min }}^{2}}\right)\left(\frac{\log n}{n}\right)+\frac{1}{(\rho-1) n}\right]$ for some absolute constant $C>0$ follows from Theorem 7 in [4]. Consequently if $\Delta_{\min }=\omega\left(\sqrt{\frac{\log n}{n}}\right)$, it would follow from Markov's inequality that $\sum_{i \in[K] \backslash I} N_{i}(n) / n=o_{p}(1)$. Thus, any arm $i \in \mathcal{I}$ must satisfy $N_{i}(n) / n \xrightarrow{p} 1 /|\mathcal{I}|$.

## A. 10 Proof of Fact 2 and Fact 3

## A.10.1 Fact 2

Since $\theta_{k}$ is $\operatorname{Beta}(1, k+1)$-distributed, its PDF, say $f_{k}(\cdot)$, is given by

$$
\begin{align*}
& \quad f_{k}(x)=(k+1)(1-x)^{k} ; \quad x \in[0,1] .  \tag{A.53}\\
& \therefore \mathbb{P}\left(\theta_{k}>\tilde{\theta}_{l}\right)=\int_{0}^{1}\left(\int_{y}^{1} f_{k}(x) d x\right) f_{l}(y) d y \\
& = \\
& =\int_{0}^{1}\left(\int_{y}^{1}(k+1)(1-x)^{k} d x\right)(l+1)(1-y)^{l} d y \quad \quad \text { (using (A.53)) } \\
& = \\
& =\int_{0}^{1}(l+1)(1-y)^{k+l+1} d y \\
& = \\
& k+l+2
\end{align*}
$$

## A.10.2 Fact 3

Since $\theta_{k}$ is $\operatorname{Beta}(k+1,1)$-distributed, its PDF, say $f_{k}(\cdot)$, is given by

$$
\begin{align*}
& f_{k}(x)=(k+1) x^{k} ; x \in[0,1]  \tag{A.54}\\
\therefore \mathbb{P}\left(\theta_{k}>\tilde{\theta}_{l}\right)= & \int_{0}^{1}\left(\int_{y}^{1} f_{k}(x) d x\right) f_{l}(y) d y \\
= & \int_{0}^{1}\left(\int_{y}^{1}(k+1) x^{k} d x\right)(l+1) y^{l} d y \quad \quad(\text { using (A.54)) } \\
= & \int_{0}^{1}(l+1)\left(1-y^{k+1}\right) y^{l} d y \\
= & 1-\frac{l+1}{k+l+2} \\
= & \frac{k+1}{k+l+2}
\end{align*}
$$

# Appendix B: Appendix to Chapter 2 

## General organization

1. §B. 1 provides the proof of Theorem 6 .
2. §B. 2 provides the proof of Theorem 7 .
3. §B. 3 provides the proof of Theorem 8.
4. §B. 4 provides the proof of Theorem 9 .
5. §B. 5 provides the proof of Proposition 1.
6. §B. 6 provides the proof of Theorem 10.
7. §B. 7 provides auxiliary results used in the analysis of ALG3.
8. §B.8 provides the proof of Theorem 11.
9. §B. 9 provides auxiliary results used in the analysis of ALG4.
10. §B. 10 provides the proof of Theorem 12.

## B. 1 Proof of Theorem 6

Notation. For each $i \in\{1,2\}$, let $\mathcal{G}_{i}(x)$ be an arbitrary collection of distributions with mean $x \in \mathbb{R}$. The tuple $\left(\mathcal{G}_{1}(x), \mathcal{G}_{2}(y)\right)$ will be referred to as an instance.

Since the horizon of play is fixed at $n$, the decision maker may play at most $n$ distinct arms. Therefore, it suffices to focus only on the sequence of the first $n$ arms that may be played. A realization of an instance $v=\left(\mathcal{G}_{1}\left(\mu_{1}\right), \mathcal{G}_{2}\left(\mu_{2}\right)\right)$ is defined as the $n$-tuple $r \equiv\left(r_{i}\right)_{1 \leqslant i \leqslant n}$, where
$r_{i} \in \mathcal{G}_{1}\left(\mu_{1}\right) \cup \mathcal{G}_{2}\left(\mu_{2}\right)$ denotes the reward distribution of arm $i \in\{1, \ldots, n\}$. It must be noted that the decision maker need not play every $\operatorname{arm} \operatorname{in} r$. Let $i^{*}:=\arg \max _{i \in\{1,2\}} \mu_{i}$. Suppose that the distribution over possible realizations of $v=\left(\mathcal{G}_{1}\left(\mu_{1}\right), \mathcal{G}_{2}\left(\mu_{2}\right)\right)$ in $\left\{r: r_{i} \in \mathcal{G}_{1}\left(\mu_{1}\right) \cup \mathcal{G}_{2}\left(\mu_{2}\right), 1 \leqslant i \leqslant n\right\}$ satisfies $\mathbb{P}\left(r_{i} \in \mathcal{G}_{i^{*}}\left(\mu_{i^{*}}\right)\right)=\alpha^{*}$ (where $\alpha^{*} \in(0,1)$ is arbitrary) for all $i \in\{1, \ldots, n\}$, i.e., optimal arms occur in the reservoir with probability $\alpha^{*}$.

Recall that the cumulative pseudo-regret after $n$ plays of a policy $\pi$ on $v=\left(\mathcal{G}_{1}\left(\mu_{1}\right), \mathcal{G}_{2}\left(\mu_{2}\right)\right)$ is given by $R_{n}^{\pi}(v)=\sum_{t=1}^{n}\left(\mu_{i^{*}}-\mu_{\mathcal{T}\left(\pi_{t}\right)}\right)$, where $\mathcal{T}\left(\pi_{t}\right) \in\{1,2\}$ indicates the type of the arm played by $\pi$ at time $t$. Our goal is to lower bound $\mathbb{E} R_{n}^{\pi}(v)$, where the expectation is w.r.t. the randomness in $\pi$ as well as the distribution over possible realizations of $v$. To this end, we define the notion of expected cumulative regret of $\pi$ on a realization $r$ of $v=\left(\mathcal{G}_{1}\left(\mu_{1}\right), \mathcal{G}_{2}\left(\mu_{2}\right)\right)$ by

$$
S_{n}^{\pi}(v, r):=\mathbb{E}^{\pi}\left[\sum_{t=1}^{n}\left(\mu_{i^{*}}-\mu_{\mathcal{T}}\left(\pi_{t}\right)\right)\right],
$$

where the expectation $\mathbb{E}^{\pi}$ is w.r.t. the randomness in $\pi$. Note that $\mathbb{E} R_{n}^{\pi}(v)=\mathbb{E}^{v} S_{n}^{\pi}(v, r)$, where the expectation $\mathbb{E}^{v}$ is w.r.t. the distribution over possible realizations of $v$. We define our problem class $\mathcal{N}_{\underline{\Delta}}$ as the collection of $\underline{\Delta}$-separated instances given by

$$
\mathcal{N}_{\underline{\Delta}}:=\left\{\left(\mathcal{G}_{1}\left(\mu_{1}\right), \mathcal{G}_{2}\left(\mu_{2}\right)\right): \mu_{1}-\mu_{2}=\underline{\Delta},\left(\mu_{1}, \mu_{2}\right) \in \mathbb{R}^{2}\right\} .
$$

Fix an arbitrary $\underline{\Delta}>0$ and consider an instance $v=\left(\left\{Q_{1}\right\},\left\{Q_{2}\right\}\right) \in \mathcal{N}_{\underline{\Delta}}$, where $\left(Q_{1}, Q_{2}\right)$ are unitvariance Gaussian distributions with means $\left(\mu_{1}, \mu_{2}\right)$ respectively. Consider an arbitrary realization $r \in\left\{Q_{1}, Q_{2}\right\}^{n}$ of $v$ and let $\mathcal{I} \subseteq\{1, \ldots, n\}$ denote the set of inferior arms in $r$ (arms with reward distribution $Q_{2}$ ). Consider another instance $v^{\prime} \in \mathcal{N}_{\Delta}$ given by $v^{\prime}=\left(\left\{\widetilde{Q}_{1}\right\},\left\{Q_{1}\right\}\right)$, where $\widetilde{Q}_{1}$ is another unit variance Gaussian with mean $\mu_{1}+\underline{\Delta}$. Now consider a realization $r^{\prime} \in\left\{\widetilde{Q}_{1}, Q_{1}\right\}^{n}$ of $v^{\prime}$ that is such that the arms at positions in $I$ have distribution $\widetilde{Q}_{1}$ while those at positions in $\{1, \ldots, n\} \backslash I$ have distribution $Q_{1}$. Notice that $\mathcal{I}$ is the set of optimal arms in $r^{\prime}$ (arms with reward
distribution $\widetilde{Q}_{1}$ ). Then, the following always holds:

$$
S_{n}^{\pi}(v, r)+S_{n}^{\pi}\left(v^{\prime}, r^{\prime}\right) \geqslant\left(\frac{\Delta n}{2}\right)\left(\mathbb{P}_{v, r}^{\pi}\left(\sum_{i \in \mathcal{I}} N_{i}(n)>\frac{n}{2}\right)+\mathbb{P}_{v^{\prime}, r^{\prime}}^{\pi}\left(\sum_{i \in \mathcal{I}} N_{i}(n) \leqslant \frac{n}{2}\right)\right),
$$

where $\mathbb{P}_{\nu, r}^{\pi}(\cdot)$ and $\mathbb{P}_{v^{\prime}, r^{\prime}}^{\pi}(\cdot)$ denote the probability measures w.r.t. the instance-realization pairs $(v, r)$ and $\left(v^{\prime}, r^{\prime}\right)$ respectively, and $N_{i}(n)$ denotes the number of plays up to and including time $n$ of arm $i \in\{1, \ldots, n\}$. Using the Bretagnolle-Huber inequality (Theorem 14.2 of [9]), we obtain

$$
S_{n}^{\pi}(v, r)+S_{n}^{\pi}\left(v^{\prime}, r^{\prime}\right) \geqslant\left(\frac{\Delta n}{4}\right) \exp \left(-\mathrm{D}\left(\mathbb{P}_{v, r}^{\pi}, \mathbb{P}_{v^{\prime}, r^{\prime}}^{\pi}\right)\right),
$$

where $\mathrm{D}\left(\mathbb{P}_{\nu, r}^{\pi}, \mathbb{P}_{\nu^{\prime}, r^{\prime}}^{\pi}\right)$ denotes the KL-Divergence between $\mathbb{P}_{v, r}^{\pi}$ and $\mathbb{P}_{\nu^{\prime}, r^{\prime}}^{\pi}$. Using Divergence decomposition (Lemma 15.1 of [9]), we further obtain

$$
S_{n}^{\pi}(v, r)+S_{n}^{\pi}\left(v^{\prime}, r^{\prime}\right) \geqslant\left(\frac{\Delta \underline{\Delta}}{4}\right) \exp \left(-\left(\frac{\mathrm{D}\left(Q_{2}, \widetilde{Q}_{1}\right)}{\underline{\Delta}}\right) S_{n}^{\pi}(v, r)\right)=\left(\frac{\underline{\Delta} n}{4}\right) \exp \left(-2 \underline{\Delta} S_{n}^{\pi}(v, r)\right),
$$

where the equality follows since $\widetilde{Q}_{1}$ and $Q_{2}$ are unit variance Gaussian distributions with means separated by $2 \Delta \underline{\Delta}$. Next, taking the expectation $\mathbb{E}^{v}$ on both sides followed by a direct application of Jensen's inequality yields

$$
\begin{equation*}
\mathbb{E} R_{n}^{\pi}(v)+\mathbb{E}^{v} S_{n}^{\pi}\left(v^{\prime}, r^{\prime}\right) \geqslant\left(\frac{\underline{\Delta} n}{4}\right) \exp \left(-2 \underline{\Delta} \mathbb{E} R_{n}^{\pi}(v)\right) \tag{B.1}
\end{equation*}
$$

Consider the $\mathbb{E}^{\nu} S_{n}^{\pi}\left(v^{\prime}, r^{\prime}\right)$ term in (B.1). Using a simple change-of-measure argument, we obtain

$$
\mathbb{E}^{v} S_{n}^{\pi}\left(v^{\prime}, r^{\prime}\right)=\mathbb{E}^{v^{\prime}}\left[S_{n}^{\pi}\left(v^{\prime}, r^{\prime}\right)\left(\frac{1-\alpha^{*}}{\alpha^{*}}\right)^{2\left(\Lambda\left(r^{\prime}\right)-n / 2\right)}\right]
$$

where $\Lambda\left(r^{\prime}\right)$ is the number of optimal arms in realization $r^{\prime}$. Since $\alpha^{*}$ is arbitrary, we fix $\alpha^{*}=1 / 2$
to obtain

$$
\begin{equation*}
\mathbb{E}^{v} S_{n}^{\pi}\left(v^{\prime}, r^{\prime}\right)=\mathbb{E}^{\nu^{\prime}} S_{n}^{\pi}\left(v^{\prime}, r^{\prime}\right)=\mathbb{E} R_{n}^{\pi}\left(v^{\prime}\right), \tag{B.2}
\end{equation*}
$$

Now, from (B.1) and (B.2), we have that for $\alpha^{*}=1 / 2$,

$$
\begin{align*}
& \mathbb{E} R_{n}^{\pi}(v)+\mathbb{E} R_{n}^{\pi}\left(v^{\prime}\right) \geqslant \frac{\Delta n}{4} \exp \left(-2 \underline{\Delta} \mathbb{E} R_{n}^{\pi}(v)\right) \\
\Longrightarrow & \tilde{R}_{n} \geqslant \frac{\Delta n}{8} \exp \left(-2 \underline{\Delta} \tilde{R}_{n}\right), \tag{B.3}
\end{align*}
$$

where $\tilde{R}_{n}:=\max \left(\mathbb{E} R_{n}^{\pi}(v), \mathbb{E} R_{n}^{\pi}\left(v^{\prime}\right)\right)$.

## Instance-dependent lower bound

The assertion of the theorem follows from the fact that the inequality (B.3) is fulfilled only if for any $\varepsilon \in(0,1), \tilde{R}_{n}$ satisfies for all $n$ large enough $\tilde{R}_{n} \geqslant(1-\varepsilon) \log n /(2 \underline{\Delta})$. Therefore, there exists an instance $v$ with gap $\underline{\Delta}$ such that $\mathbb{E} R_{n}^{\pi}(v) \geqslant C \log n / \underline{\Delta}$ for some absolute constant $C$ and $n$ large enough, whenever $\alpha^{*}=1 / 2$. In fact, said statement holds for all $\alpha^{*} \leqslant 1 / 2$ since the policy $\pi$ satisfies Definition 1.

## Instance-independent (minimax) lower bound

Since $\tilde{R}_{n} \leqslant \underline{\Delta} n$, it follows from (B.3) that

$$
\tilde{R}_{n} \geqslant \frac{\Delta n}{8} \exp \left(-2 \underline{\Delta}^{2} n\right) .
$$

Setting $\underline{\Delta}=1 / \sqrt{n}$, and noting that the inequality, in fact, holds for all $\alpha^{*} \leqslant 1 / 2$ (owing to the admissibility of $\pi$; see Definition 1), proves the stated assertion.

## B. 2 Proof of Theorem 7

Note that this result is stated for general $K \geqslant 2$ and is not specific to $K=2$. In fact, the nature of the set of possible sub-optimal types is inconsequential to the proof that follows as long as said set is at least $\underline{\Delta}$-separated from the optimal mean reward. Consider an arbitrary policy $\pi \in \tilde{\Pi}$. Denote
by $A_{n}^{\pi}$ the number of distinct arms played by $\pi$ until time $n$. Consider an arbitrary $k \in\{1, \ldots, n\}$. Then, conditioned on $A_{n}^{\pi}=k$, the expected cumulative regret incurred by $\pi$ is at least

$$
\begin{equation*}
\mathbb{E}\left[R_{n}^{\pi} \mid A_{n}^{\pi}=k\right] \geqslant\left(1-\alpha_{1}\right) \underline{\Delta} k+\left(1-\alpha_{1}\right)^{k} \underline{\Delta}(n-k)=: f(k) . \tag{B.4}
\end{equation*}
$$

Intuition behind (B.4). Each of the $k$ arms played during the horizon has at least one pull associated with it. Consider a clairvoyant policy coupled to $\pi$ that learns the best among the $A_{n}^{\pi}$ arms played by $\pi$ as soon as each has been pulled exactly once, i.e., after a total of $A_{n}^{\pi}$ pulls. Clearly, the regret incurred by said clairvoyant policy lower bounds $\mathbb{E} R_{n}^{\pi}$. Further, since $A_{n}^{\pi}$ is independent of the sample-history of arms, it follows that the $A_{n}^{\pi}$ arms are statistically identical. Thus, conditioned on $A_{n}^{\pi}=k$, the expected regret from the first $k$ pulls of the clairvoyant policy is at least $\left(1-\alpha_{1}\right) \underline{\Delta} k$. Also, the probability that each of the $k$ arms is inferior-typed is $\left(1-\alpha_{1}\right)^{k}$; the clairvoyant policy thus incurs a regret of at least $\left(1-\alpha_{1}\right)^{k} \underline{\Delta}(n-k)$ going forward. This explains the lower bound in (B.4). Therefore, for any $k \in\{1,2, \ldots, n\}$, we have

$$
\mathbb{E}\left[R_{n}^{\pi} \mid A_{n}^{\pi}=k\right] \geqslant \min _{k \in\{1,2, \ldots, n\}} f(k) \geqslant \min _{x \in[0, n]} f(x) .
$$

We will show that $f(x)$ is strictly convex over $[0, n]$ with $f^{\prime}(0)<0$ and $f^{\prime}(n)>0$. Then, it would follow that $f(\cdot)$ admits a unique minimizer $x_{n}^{*} \in(0, n)$ given by the solution to $f^{\prime}(x)=0$. The minimum $f\left(x_{n}^{*}\right)$ will turn out to be logarithmic in $n$. Observe that

$$
\begin{aligned}
& f^{\prime}(x)=\left(1-\alpha_{1}\right) \underline{\Delta}+\left(1-\alpha_{1}\right)^{x} \underline{\Delta}\left[(n-x) \log \left(1-\alpha_{1}\right)-1\right] \\
& f^{\prime \prime}(x)=-\left(1-\alpha_{1}\right)^{x} \underline{\Delta}\left[2-(n-x) \log \left(1-\alpha_{1}\right)\right] \log \left(1-\alpha_{1}\right)
\end{aligned}
$$

Since $\underline{\Delta}>0$, it follows that $f^{\prime \prime}(x)>0$ over $[0, n]$. Further, note that

$$
\begin{aligned}
& f^{\prime}(0)=-\alpha_{1} \underline{\Delta}+\underline{\Delta} n \log \left(1-\alpha_{1}\right)<0 \\
& f^{\prime}(n)=\left(1-\alpha_{1}\right) \underline{\Delta}-\left(1-\alpha_{1}\right)^{n} \underline{\Delta}>0
\end{aligned}
$$

Solving $f^{\prime}\left(x_{n}^{*}\right)=0$ for the unique minimizer $x_{n}^{*}$, we obtain

$$
\begin{aligned}
& \left(\frac{1}{1-\alpha_{1}}\right)^{x_{n}^{*}-1}-1=\left(n-x_{n}^{*}\right) \log \left(\frac{1}{1-\alpha_{1}}\right) \\
\Longrightarrow & \left(\frac{1}{1-\alpha_{1}}\right)^{x_{n}^{*}}+x_{n}^{*} \log \left(\frac{1}{1-\alpha_{1}}\right)>n \log \left(\frac{1}{1-\alpha_{1}}\right) \\
\Longrightarrow & 2\left(\frac{1}{1-\alpha_{1}}\right)^{x_{n}^{*}}>n \log \left(\frac{1}{1-\alpha_{1}}\right),
\end{aligned}
$$

where the last inequality follows using $y>\log y$. Therefore, we have

$$
\begin{aligned}
& \left(\frac{1}{1-\alpha_{1}}\right)^{x_{n}^{*}}>\frac{n}{2} \log \left(\frac{1}{1-\alpha_{1}}\right) \\
\Longrightarrow & x_{n}^{*}>\frac{\log n+\log \log \left(\frac{1}{1-\alpha_{1}}\right)-\log 2}{\log \left(\frac{1}{1-\alpha_{1}}\right)} .
\end{aligned}
$$

Thus, for any $k \in\{1, \ldots, n\}$,

$$
\begin{aligned}
& \mathbb{E}\left[R_{n}^{\pi} \mid A_{n}^{\pi}=k\right] \geqslant f\left(x_{n}^{*}\right)>\left(1-\alpha_{1}\right) \underline{\Delta} x_{n}^{*}>\left(1-\alpha_{1}\right)\left(\frac{\log n+\log \log \left(\frac{1}{1-\alpha_{1}}\right)-\log 2}{\log \left(\frac{1}{1-\alpha_{1}}\right)}\right) \underline{\Delta} \\
\Longrightarrow & \mathbb{E} R_{n}^{\pi} \geqslant\left(1-\alpha_{1}\right)\left(\frac{\log n+\log \log \left(\frac{1}{1-\alpha_{1}}\right)-\log 2}{\log \left(\frac{1}{1-\alpha_{1}}\right)}\right) \underline{\Delta} \\
\Longrightarrow & \inf _{\pi \in \tilde{\Pi}} \frac{\mathbb{E} R_{n}^{\pi}}{\log n} \geqslant\left(1-\alpha_{1}\right)\left(\frac{1}{\log \left(\frac{1}{1-\alpha_{1}}\right)}+\frac{\log \log \left(\frac{1}{1-\alpha_{1}}\right)-\log 2}{(\log n) \log \left(\frac{1}{1-\alpha_{1}}\right)}\right) \underline{\Delta} \\
\Longrightarrow & \inf _{\pi \in \tilde{\Pi}} \frac{\mathbb{E} R_{n}^{\pi}}{\log n} \geqslant\left(1-\alpha_{1}\right)\left(\frac{1-\alpha_{1}}{\alpha_{1}}+\frac{\log \log \left(\frac{1}{1-\alpha_{1}}\right)-\log 2}{(\log n) \log \left(\frac{1}{1-\alpha_{1}}\right)}\right) \underline{\Delta} \\
\Longrightarrow & \inf _{\pi \in \tilde{\Pi}} \frac{\mathbb{E} R_{n}^{\pi}}{\log n} \geqslant \frac{\left(1-\alpha_{1}\right)^{2} \underline{\Delta}}{\alpha_{1}}+\left(1-\alpha_{1}\right)\left(\frac{\log \log \left(\frac{1}{1-\alpha_{1}}\right)-\log 2}{(\log n) \log \left(\frac{1}{1-\alpha_{1}}\right)}\right) \underline{\Delta},
\end{aligned}
$$

where $(\dagger)$ follows using $\log y \leqslant y-1$. Taking the appropriate limit now proves the assertion.

## B. 3 Proof of Theorem 8

The reservoir distribution is given by $\alpha=\left(\alpha_{1}, \ldots, \alpha_{K}\right)$. In the full information setting, the decision maker observes the true mean reward of an arm immediately upon pulling it. Let $\pi=$ $\left(\pi_{t}: t=1,2, \ldots\right)$ be the policy that pulls a new arm from the reservoir in each period. Let $N$ denote the first time at which one arm of each of the $K$ types is collected under $\pi$. Then, it follows from classical results (see Theorem 4.1 in [71]) for the Coupon-collector problem that

$$
\begin{align*}
& \mathbb{E} N=\int_{0}^{\infty}\left(1-\prod_{j=1}^{K}\left(1-\exp \left(-\alpha_{j} y\right)\right)\right) d y \underset{(\dagger)}{\geqslant} \int_{0}^{\infty}\left(1-\prod_{j=1}^{K}\left(1-\exp \left(-\frac{y}{K}\right)\right)\right) d y \\
& \text { (ま) } \sum_{j=1}^{K} \frac{K}{j} \geqslant K \log K \text {, } \tag{B.5}
\end{align*}
$$

where $(\dagger)$ follows as $\prod_{j=1}^{K}\left(1-\exp \left(-\alpha_{j} y\right)\right)$ is maximized when $\alpha$ is the Uniform distribution; $(\ddagger)$ is a classical result (see previous reference). The optimal policy $\pi^{*}$ follows $\pi$ until time $N$, and subsequently commits to the arm with the highest mean among the first $N$ arms. The lifetime regret of $\pi^{*}$ is then given by

$$
\begin{align*}
\mathbb{E} R_{\infty}^{\pi^{*}}=\mathbb{E}\left[\sum_{t=1}^{N} \sum_{i=2}^{K}\left(\mu_{1}-\mu_{i}\right) \mathbb{1}\left\{\mathcal{T}\left(\pi_{t}\right)=i\right\}\right] & =\mathbb{E}\left[\sum_{t=1}^{\infty} \sum_{i=2}^{K}\left(\mu_{1}-\mu_{i}\right) \mathbb{1}\left\{\mathcal{T}\left(\pi_{t}\right)=i, t \leqslant N\right\}\right] \\
& =\sum_{t=1}^{\infty} \sum_{i=2}^{K}\left(\mu_{1}-\mu_{i}\right) \mathbb{P}\left(\mathcal{T}\left(\pi_{t}\right)=i, t \leqslant N\right), \tag{B.6}
\end{align*}
$$

where the last equality follows from Tonelli's Theorem. Note that

$$
\begin{align*}
\mathbb{P}\left(\mathcal{T}\left(\pi_{t}\right)=i, t \leqslant N\right) & =\mathbb{P}\left(\mathcal{T}\left(\pi_{t}\right)=i\right)-\mathbb{P}\left(\mathcal{T}\left(\pi_{t}\right)=i, t>N\right) \\
& =\alpha_{i}-\mathbb{P}\left(\mathcal{T}\left(\pi_{t}\right)=i \mid t>N\right) \mathbb{P}(t>N) \\
& =\alpha_{(\star)}-\mathbb{P}\left(\mathcal{T}\left(\pi_{t}\right)=i\right) \mathbb{P}(t>N) \\
& =\alpha_{i}-\alpha_{i} \mathbb{P}(t>N) \\
& =\alpha_{i} \mathbb{P}(N \geqslant t), \tag{B.7}
\end{align*}
$$

where $(\star)$ follows since $\mathcal{T}\left(\pi_{t}\right)$ is i.i.d. in time $t$. Therefore, from (B.6) and (B.7), we have

$$
\begin{align*}
\mathbb{E} R_{\infty}^{\pi^{*}} & =\sum_{t=1}^{\infty} \sum_{i=2}^{K} \alpha_{i}\left(\mu_{1}-\mu_{i}\right) \mathbb{P}(N \geqslant t) \\
& =\sum_{i=2}^{K} \alpha_{i}\left(\mu_{1}-\mu_{i}\right) \sum_{t=1}^{\infty} \mathbb{P}(N \geqslant t) \\
& =\sum_{i=2}^{K} \alpha_{i}\left(\mu_{1}-\mu_{i}\right) \mathbb{E} N . \tag{B.8}
\end{align*}
$$

Finally, from (B.8) and (B.5), one obtains

$$
\mathbb{E} R_{\infty}^{\pi^{*}} \geqslant \sum_{i=2}^{K} \alpha_{i}\left(\mu_{1}-\mu_{i}\right) K \log K
$$

## B. 4 Proof of Theorem 9

Let $\left\{\left(X_{1, j}^{l}, \ldots, X_{K, j}^{l}\right): j=1, \ldots, m_{l}\right\}$ be the reward sequence associated with the $K$ arms played in the $l^{\text {th }}$ epoch, where $m_{l}=\left\lceil e^{2 \sqrt{l}} \log n\right\rceil$. Let $\mathcal{A}:=\{1, \ldots, K\}$ and define

$$
I:=\inf \left\{l \in \mathbb{N}:\left|\sum_{j=1}^{m_{l}}\left(X_{a, j}^{l}-X_{b, j}^{l}\right)\right| \geqslant 2 m_{l} e^{-\sqrt{l}} \forall a, b \in \mathcal{A}, a<b\right\} .
$$

Then,

$$
\begin{aligned}
\mathbb{P}(I \geqslant k) & =\mathbb{P}\left(\bigcap_{l=1}^{k-1} \bigcup_{a, b \in \mathcal{A}, a<b}\left\{\left|\sum_{j=1}^{m_{l}}\left(X_{a, j}^{l}-X_{b, j}^{l}\right)\right|<2 m_{l} e^{-\sqrt{l}}\right\}\right) \\
& =\prod_{l=1}^{k-1} \mathbb{P}\left(\bigcup_{a, b \in \mathcal{A}, a<b}\left\{\left|\sum_{j=1}^{m_{l}}\left(X_{a, j}^{l}-X_{b, j}^{l}\right)\right|<2 m_{l} e^{-\sqrt{l}}\right\}\right) \\
& \leqslant \prod_{l=1}^{k-1}\left[1-\mathbb{P}(\mathrm{D})+\mathbb{P}(\mathrm{D}) \mathbb{P}_{\mathrm{D}}\left(\bigcup_{a, b \in \mathcal{A}, a<b}\left\{\left|\sum_{j=1}^{m_{l}}\left(X_{a, j}^{l}-X_{b, j}^{l}\right)\right|<2 m_{l} e^{-\sqrt{l}}\right\}\right)\right],
\end{aligned}
$$

where D denotes the event that the $K$ arms played in epoch $l$ are "all-distinct," i.e., no two arms belong to the same type, $\mathbb{P}_{\mathrm{D}}(\cdot):=\mathbb{P}(\cdot \mid \mathrm{D})$ denotes the corresponding conditional measure, and $\mathbb{P}(\mathrm{D})=K!\prod_{i=1}^{K} \alpha_{i}$. Using the Union bound, we obtain

$$
\begin{equation*}
\mathbb{P}(I \geqslant k) \leqslant \prod_{l=1}^{k-1}\left[1-\mathbb{P}(\mathrm{D})+\mathbb{P}(\mathrm{D}) \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}_{\mathrm{D}}\left(\left|\sum_{j=1}^{m_{l}}\left(X_{a, j}^{l}-X_{b, j}^{l}\right)\right|<2 m_{l} e^{-\sqrt{l}}\right)\right] . \tag{B.9}
\end{equation*}
$$

Define $\tau:=K \sum_{l=1}^{I} m_{l}$. Consider the following events:

- $\mathrm{E}_{1}:=\{$ None of the arms played in epoch $I$ belongs to the optimal type $\}$.
- $E_{2}:=\{$ At least one optimal-typed arm is played in epoch $I$, and the empirically best arm is not optimal-typed $\}$.

Recall that $\bar{\Delta}=\mu_{1}-\mu_{K}$ denotes the maximal sub-optimality gap. Then, the cumulative pseudoregret $R_{n}$ (superscript $\pi$ suppressed for notational convenience) of ALG1 $(n)$ is bounded as

$$
\begin{aligned}
R_{n} & \leqslant \mathbb{1}\{\tau \leqslant n\}\left[\bar{\Delta} \tau+\mathbb{1}\left\{\mathrm{E}_{1} \cup \mathrm{E}_{2}\right\} \bar{\Delta} n\right]+\mathbb{1}\{\tau>n\} \bar{\Delta} n \\
& \leqslant \bar{\Delta} \tau+\left(\mathbb{1}\left\{\tau \leqslant n, \mathrm{E}_{1}\right\}+\mathbb{1}\left\{\tau \leqslant n, \mathrm{E}_{2}\right\}\right) \bar{\Delta} n+\mathbb{1}\{\tau>n\} \bar{\Delta} n .
\end{aligned}
$$

Taking expectations,

$$
\begin{aligned}
\mathbb{E} R_{n} & \leqslant \bar{\Delta} \mathbb{E} \tau+\left[\mathbb{P}\left(\tau \leqslant n, \mathrm{E}_{1}\right)+\mathbb{P}\left(\tau \leqslant n, \mathrm{E}_{2}\right)\right] \bar{\Delta} n+\mathbb{P}(\tau>n) \bar{\Delta} n \\
& \leqslant 2 \bar{\Delta} \mathbb{E} \tau+\left[\mathbb{P}\left(\tau \leqslant n, \mathrm{E}_{1}\right)+\mathbb{P}\left(\tau \leqslant n, \mathrm{E}_{2}\right)\right] \bar{\Delta} n
\end{aligned}
$$

where the last step uses Markov's inequality. Therefore,

$$
\begin{aligned}
\mathbb{E} R_{n} & \leqslant 2 K \bar{\Delta} \mathbb{E}\left[\operatorname{Im}_{I}\right]+\left[\mathbb{P}\left(\tau \leqslant n, \mathrm{E}_{1}\right)+\mathbb{P}\left(\tau \leqslant n, \mathrm{E}_{2}\right)\right] \bar{\Delta} n \\
& \leqslant 4 K \bar{\Delta} \mathbb{E}\left[I e^{2 \sqrt{I}}\right] \log n+\left[\mathbb{P}\left(\tau \leqslant n, \mathrm{E}_{1}\right)+\mathbb{P}\left(\tau \leqslant n, \mathrm{E}_{2}\right)\right] \bar{\Delta} n .
\end{aligned}
$$

Upper bounding $\mathbb{E}\left[I e^{2 \sqrt{I}}\right]$
Recall that $\delta=\min _{1 \leqslant i<j \leqslant K}\left(\mu_{i}-\mu_{j}\right)$ denotes the smallest gap between any two distinct mean rewards. Then, on the event D , for any $a, b \in \mathcal{A}, a<b$, we either have $\mathbb{E}\left[X_{a, j}^{l}-X_{b, j}^{l}\right] \geqslant \delta$ or $\mathbb{E}\left[X_{a, j}^{l}-X_{b, j}^{l}\right] \leqslant-\delta$. Without loss of generality, suppose that $\mathbb{E}\left[X_{a, j}^{l}-X_{b, j}^{l}\right] \geqslant \delta$. Then,

$$
\begin{aligned}
\mathbb{P}_{\mathrm{D}}\left(\left|\sum_{j=1}^{m_{l}}\left(X_{a, j}^{l}-X_{b, j}^{l}\right)\right|<2 m_{l} e^{-\sqrt{l}}\right) & \leqslant \mathbb{P}_{\mathrm{D}}\left(\sum_{j=1}^{m_{l}}\left(X_{a, j}^{l}-X_{b, j}^{l}\right)<2 m_{l} e^{-\sqrt{l}}\right) \\
& =\mathbb{P}_{\mathrm{D}}\left(\sum_{j=1}^{m_{l}}\left(X_{a, j}^{l}-X_{b, j}^{l}-\delta\right)<-m_{l}\left(\delta-2 e^{-\sqrt{l}}\right)\right) .
\end{aligned}
$$

Then, for $l>\left\lceil\log ^{2}(4 / \delta)\right\rceil=: k^{*}$, one has that

$$
\begin{equation*}
\mathbb{P}_{\mathrm{D}}\left(\left|\sum_{j=1}^{m_{l}}\left(X_{a, j}^{l}-X_{b, j}^{l}\right)\right|<2 m_{l} e^{-\sqrt{l}}\right) \leqslant \mathbb{P}_{\mathrm{D}}\left(\sum_{j=1}^{m_{l}}\left(X_{a, j}^{l}-X_{b, j}^{l}-\delta\right)<-2 m_{l} e^{-\sqrt{l}}\right) \leqslant n^{-2}, \tag{B.10}
\end{equation*}
$$

where the final inequality follows using the Chernoff-Hoeffding bound [69], together with the fact that $-1 \leqslant X_{a, j}^{l}-X_{b, j}^{l} \leqslant 1$ and $m_{l}=\left\lceil e^{2 \sqrt{l}} \log n\right\rceil$. Using (B.9) and (B.10), we obtain for $k>k^{*}+1$
that

$$
\begin{aligned}
\mathbb{P}(I \geqslant k) & \leqslant \prod_{l=1}^{k-1}\left[1-\mathbb{P}(\mathrm{D})+\mathbb{P}(\mathrm{D}) \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}_{\mathrm{D}}\left(\left|\sum_{j=1}^{m_{l}}\left(X_{a, j}^{l}-X_{b, j}^{l}\right)\right|<2 m_{l} e^{-\sqrt{l}}\right)\right] \\
& \leqslant \prod_{k^{*}<l \leqslant k-1}\left[1-\mathbb{P}(\mathrm{D})+\frac{K^{2} \mathbb{P}(\mathrm{D})}{n^{2}}\right] \\
& \leqslant\left[1-\frac{3 \mathbb{P}(\mathrm{D})}{4}\right]^{k-k^{*}-1},
\end{aligned}
$$

where the last inequality holds for $n \geqslant 2 K$ (We will ensure that all guarantees hold for $n \geqslant K$ by offsetting regret by $2 K \bar{\Delta}$ in the end). Thus, for any $k \geqslant 1$ and $n \geqslant 2 K$, we have

$$
\mathbb{P}\left(I \geqslant k^{*}+k\right) \leqslant\left[1-\frac{3 \mathbb{P}(\mathrm{D})}{4}\right]^{k-1}
$$

Now,

$$
\begin{aligned}
\mathbb{E}\left[I e^{2 \sqrt{I}}\right] & \leqslant k^{*} e^{2 \sqrt{k^{*}}}+\sum_{k=1}^{\infty}\left(k^{*}+k\right) e^{2 \sqrt{k^{*}+k}} \mathbb{P}\left(I=k^{*}+k\right) \\
& \leqslant k^{*} e^{2 \sqrt{k^{*}}}+\sum_{k=1}^{\infty}\left(k^{*}+k\right) e^{2 \sqrt{k^{*}}} e^{2 \sqrt{k}} \mathbb{P}\left(I=k^{*}+k\right) \\
& \leqslant k^{*} e^{2 \sqrt{k^{*}}}+2 \sum_{k=1}^{\infty} k^{*} k e^{2 \sqrt{k^{*}}} e^{2 \sqrt{k}} \mathbb{P}\left(I=k^{*}+k\right) \\
& \leqslant k^{*} e^{2 \sqrt{k^{*}}}+2 \sum_{k=1}^{\infty} k^{*} k e^{2 \sqrt{k^{*}}} e^{2 \sqrt{k}} \mathbb{P}\left(I \geqslant k^{*}+k\right) \\
& \leqslant k^{*} e^{2 \sqrt{k^{*}}}+2 \sum_{k=1}^{\infty} k^{*} k e^{2 \sqrt{k^{*}}} e^{2 \sqrt{k}}\left[1-\frac{3 \mathbb{P}(\mathrm{D})}{4}\right]^{k-1} \\
& =k^{*} e^{2 \sqrt{k^{*}}}\left[1+2 \sum_{k=1}^{\infty} k e^{2 \sqrt{k}}\left[1-\frac{3 \mathbb{P}(\mathrm{D})}{4}\right]^{k-1}\right] \\
& \leqslant 4 k^{*} e^{2 \sqrt{k^{*}}} \sum_{k=1}^{\infty} k e^{2 \sqrt{k}}\left[1-\frac{3 \mathbb{P}(\mathrm{D})}{4}\right]^{k-1},
\end{aligned}
$$

where $(\dagger)$ follows using $\sqrt{k^{*}+k} \leqslant \sqrt{k^{*}}+\sqrt{k}$, and ( $\ddagger$ ) using $k^{*}+k \leqslant 2 k^{*} k$ (since $k^{*}, k \in \mathbb{N}$ ). Define

$$
C_{\alpha}:=\sum_{k=1}^{\infty} k e^{2 \sqrt{k}}\left[1-\frac{3 \mathbb{P}(\mathrm{D})}{4}\right]^{k-1} .
$$

Since $\mathbb{P}(\mathrm{D})=K!\prod_{i=1}^{K} \alpha_{i}$, note that the infinite summation is finite since $\alpha=\left(\alpha_{i}: i=1, \ldots, K\right)$ is coordinate-wise bounded away from 0 . Therefore,

$$
\mathbb{E}\left[I e^{2 \sqrt{I}}\right] \leqslant 4 C_{\alpha} k^{*} e^{2 \sqrt{k^{*}}} .
$$

Note that

$$
e^{2 \sqrt{k^{*}}}=e^{2 \sqrt{\left[\log ^{2}(4 / \delta)\right\rceil}} \leqslant e^{2 \sqrt{\log ^{2}(4 / \delta)+1}} \leqslant e^{2\left(\sqrt{\log ^{2}(4 / \delta)+1}\right)} \leqslant \frac{16 e^{2}}{\delta^{2}} .
$$

Therefore, in conclusion,

$$
\mathbb{E}\left[I e^{2 \sqrt{I}}\right] \leqslant \frac{128 e^{2} C_{\alpha}}{\delta^{2}} \log ^{2}\left(\frac{4}{\delta}\right) .
$$

## Upper bounding $\mathbb{P}\left(\tau \leqslant n, \mathrm{E}_{1}\right)$

Note that on the event $\{\tau \leqslant n\}$, the duration of epoch $I$ is $K m_{I}$. On the event $\mathrm{E}_{1}$, the consideration set contains at least two arms that belong to the same type. Without loss of generality, suppose
that these arms are indexed 1 and 2 . Then,

$$
\begin{aligned}
\mathbb{P}\left(\tau \leqslant n, \mathrm{E}_{1}\right) & \leqslant \mathbb{P}\left(\left|\sum_{j=1}^{m_{I}}\left(X_{1, j}^{I}-X_{2, j}^{I}\right)\right| \geqslant 2 m_{I} e^{-\sqrt{I}}, \tau \leqslant n\right) \\
& \leqslant \mathbb{P}\left(\left|\sum_{j=1}^{m_{I}}\left(X_{1, j}^{I}-X_{2, j}^{I}\right)\right| \geqslant 2 m_{I} e^{-\sqrt{I}}, I \leqslant n\right) \\
& \leqslant \sum_{l=1}^{n} \mathbb{P}\left(\left|\sum_{j=1}^{m_{l}}\left(X_{1, j}^{l}-X_{2, j}^{l}\right)\right| \geqslant 2 m_{l} e^{-\sqrt{l}}\right) \\
& \leqslant \sum_{l=1}^{n} \frac{2}{n^{2}} \\
& =\frac{2}{n},
\end{aligned}
$$

where the last inequality follows using the Chernoff-Hoeffding bound [69].

## Upper bounding $\mathbb{P}\left(\tau \leqslant n, \mathrm{E}_{2}\right)$

Note that on the event $\{\tau \leqslant n\}$, the duration of epoch $I$ is $K m_{I}$. On the event $\mathrm{E}_{2}$, the consideration set $\mathcal{A}$ contains at least one arm of the optimal type, and the empirically best arm belongs to an inferior type. Without loss of generality, suppose that arm 1 belongs to the optimal type and $\mathcal{I} \subset \mathcal{A}$ denotes the set of inferior-typed arms. We then have

$$
\begin{aligned}
\mathbb{P}\left(\tau \leqslant n, \mathrm{E}_{2}\right) & \leqslant \mathbb{P}\left(\bigcup_{a \in \mathcal{I}}\left\{\sum_{j=1}^{m_{I}}\left(X_{a, j}^{I}-X_{1, j}^{I}\right) \geqslant 2 m_{I} e^{-\sqrt{I}}, \tau \leqslant n\right\}\right) \\
& \leqslant \sum_{a \in \mathcal{I}} \mathbb{P}\left(\sum_{j=1}^{m_{I}}\left(X_{a, j}^{I}-X_{1, j}^{I}\right) \geqslant 2 m_{I} e^{-\sqrt{I}}, I \leqslant n\right) \\
& \leqslant \sum_{a \in I} \sum_{l=1}^{n} \mathbb{P}\left(\sum_{j=1}^{m_{l}}\left(X_{a, j}^{l}-X_{1, j}^{l}\right) \geqslant 2 m_{l} e^{-\sqrt{l}}\right) \\
& \leqslant \sum_{a \in I} \sum_{l=1}^{n} \frac{1}{n^{2}} \\
& \leqslant \frac{K}{n}
\end{aligned}
$$

where the second-to-last inequality follows using Hoeffding's bound [69] since $\mathbb{E}\left[X_{a, j}^{l}-X_{1, j}^{l}\right] \leqslant$ $-\underline{\Delta}<0 \forall a \in \mathcal{I}$.

Putting everything together

In conclusion, the expected cumulative regret of the policy $\pi$ given by $\operatorname{ALG1}(n)$ is bounded for any $n \geqslant K$ as

$$
\mathbb{E} R_{n}^{\pi} \leqslant \frac{\tilde{C}_{\alpha} K \bar{\Delta} \log n}{\delta^{2}} \log ^{2}\left(\frac{4}{\delta}\right)+4 K \bar{\Delta},
$$

where $\tilde{C}_{\boldsymbol{\alpha}}$ is a finite constant that depends only on $\alpha=\left(\alpha_{i}: i=1, \ldots, K\right)$. In particular, $\tilde{C}_{\boldsymbol{\alpha}}$ is given by the following infinite summation:

$$
\begin{equation*}
\tilde{C}_{\alpha}:=512 e^{2} C_{\alpha}=512 e^{2} \sum_{k=1}^{\infty} k e^{2 \sqrt{k}}\left[1-\frac{3 K!\prod_{i=1}^{K} \alpha_{i}}{4}\right]^{k-1} . \tag{B.11}
\end{equation*}
$$

## B. 5 Proof of Proposition 1

Consider the following stopping time:

$$
\tau:=\inf \left\{m \in \mathbb{N}: \exists a, b \in \mathcal{A}, a<b \text { s.t. } \mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)<4 \sqrt{m \log m}\right\} .
$$

Since $\mathbb{P}\left(\bigcap_{m \geqslant 1} \bigcap_{a, b \in \mathcal{A}, a<b}\left\{\left|\mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)\right| \geqslant 4 \sqrt{m \log m}\right\}\right) \geqslant \mathbb{P}(\tau=\infty)$, it suffices to show that $\mathbb{P}(\tau=\infty)$ is bounded away from 0 . To this end, define the following entities:

$$
\begin{aligned}
& \Lambda_{K}:=\inf \left\{p \in \mathbb{N}: \sum_{m=p}^{\infty} \frac{1}{m^{8}} \leqslant \frac{1}{2 K^{2}}\right\}, \\
& T_{0}:=\max \left(\left[\left(\frac{64}{\delta^{2}}\right) \log ^{2}\left(\frac{64}{\delta^{2}}\right)\right], \Lambda_{K}\right),
\end{aligned}
$$

$$
f(x):=x+4 \sqrt{x \log x} \quad \text { for } x \geqslant 1 .
$$

Lemma 1 For any $a, b \in \mathcal{A}$ s.t. $a<b$, it is the case that

$$
\left\{\mathcal{Z}_{a, b}>f\left(T_{0}\right)\right\} \subseteq \bigcap_{m=1}^{T_{0}}\left\{\mathcal{Z}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right) \geqslant 4 \sqrt{m \log m}\right\}
$$

Proof of Lemma 1. Note that

$$
\begin{aligned}
\mathcal{Z}_{a, b} & >f\left(T_{0}\right) \\
& =T_{0}+4 \sqrt{T_{0} \log T_{0}} \\
& \geqslant m+4 \sqrt{m \log m} \forall 1 \leqslant m \leqslant T_{0} \\
& \geqslant \sum_{j=1}^{m}\left(X_{b, j}-X_{a, j}\right)+4 \sqrt{m \log m} \forall 1 \leqslant m \leqslant T_{0} \\
\Longrightarrow \mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right) & \geqslant 4 \sqrt{m \log m} \forall 1 \leqslant m \leqslant T_{0},
\end{aligned}
$$

where $(\mathfrak{a})$ follows since the rewards are bounded in $[0,1]$, i.e., $\left|X_{a, j}-X_{b, j}\right| \leqslant 1$.

Lemma 2 For $m \geqslant T_{0}$, it is the case that

$$
\delta \geqslant 8 \sqrt{\frac{\log m}{m}}
$$

Proof of Lemma 2. First of all, note that $T_{0} \geqslant\left(64 / \delta^{2}\right) \log ^{2}\left(64 / \delta^{2}\right) \geqslant 64$ (since $\delta \leqslant 1$ ). For
$s=\left(64 / \delta^{2}\right) \log ^{2}\left(64 / \delta^{2}\right)$, one has

$$
\delta^{2}=\frac{64 \log ^{2}\left(\frac{64}{\delta^{2}}\right)}{s} \underset{(\mathfrak{b})}{\geqslant} \frac{64\left[\log \left(\frac{64}{\delta^{2}}\right)+2 \log \log \left(\frac{64}{\delta^{2}}\right)\right]}{s}=\frac{64 \log s}{s}
$$

where (b) follows since the function $g(x):=x^{2}-x-2 \log x$ is monotone increasing for $x \geqslant \log 64$ (think of $\log \left(64 / \delta^{2}\right)$ as $x$ ), and therefore attains its minimum at $x=\log 64$; one can verify that this minimum is strictly positive. Furthermore, since $\log s / s$ is monotone decreasing for $s \geqslant 64$, it follows that for any $m \geqslant T_{0}$,

$$
\delta^{2} \geqslant \frac{64 \log m}{m}
$$

Now coming back to the proof of Proposition 1 , consider an arbitrary $l \in \mathbb{N}$ such that $l>T_{0}$. Then,

$$
\begin{aligned}
\mathbb{P}(\tau \leqslant l) & =\mathbb{P}\left(\tau \leqslant l, \mathcal{Z}>f\left(T_{0}\right)\right)+\mathbb{P}\left(\tau \leqslant l, \mathcal{Z} \leqslant f\left(T_{0}\right)\right) \\
& \leqslant \mathbb{P}\left(\tau \leqslant l, \mathcal{Z}>f\left(T_{0}\right)\right)+\Phi\left(f\left(T_{0}\right)\right)
\end{aligned}
$$

Now,

$$
\begin{aligned}
& \mathbb{P}\left(\tau \leqslant l, \mathcal{Z}>f\left(T_{0}\right)\right) \\
& =\mathbb{P}\left(\bigcup_{m=1}^{l} \bigcup_{a, b \in \mathcal{A}, a<b}\left\{\mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)<4 \sqrt{m \log m}, \mathcal{Z}_{a, b}>f\left(T_{0}\right)\right\}\right) \\
& \underset{(\dagger)}{=} \mathbb{P}\left(\bigcup_{m=T_{0}}^{l} \bigcup_{a, b \in \mathcal{A}, a<b}\left\{\mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)<4 \sqrt{m \log m}, \mathcal{Z}_{a, b}>f\left(T_{0}\right)\right\}\right) \\
& \leqslant \sum_{m=T_{0}}^{l} \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}\left(\mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)<4 \sqrt{m \log m}, \mathcal{Z}_{a, b}>f\left(T_{0}\right)\right) \\
& =\sum_{m=T_{0}}^{l} \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}\left(\mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}-\delta\right)<-m\left(\delta-4 \sqrt{\frac{\log m}{m}}\right), \mathcal{Z}_{a, b}>f\left(T_{0}\right)\right) \\
& \leqslant \sum_{m=T_{0}}^{l} \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}\left(\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}-\delta\right)<-m\left(\delta-4 \sqrt{\frac{\log m}{m}}\right), \mathcal{Z}_{a, b}>f\left(T_{0}\right)\right) \\
& \underset{\text { (京 }}{\leqslant} \sum_{m=T_{0}}^{l} \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}\left(\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}-\delta\right)<-4 \sqrt{m \log m}, \mathcal{Z}_{a, b}>f\left(T_{0}\right)\right) \\
& =\bar{\Phi}\left(f\left(T_{0}\right)\right) \sum_{m=T_{0}}^{l} \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}\left(\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}-\delta\right)<-4 \sqrt{m \log m}\right) \\
& \underset{(\star)}{\leqslant} \bar{\Phi}\left(f\left(T_{0}\right)\right) \sum_{m=T_{0}}^{l} \sum_{a, b \in \mathcal{A}, a<b} \frac{1}{m^{8}} \\
& \leqslant \bar{\Phi}\left(f\left(T_{0}\right)\right) K^{2} \sum_{m=T_{0}}^{\infty} \frac{1}{m^{8}} \\
& \underset{(*)}{\leqslant} \frac{\bar{\Phi}\left(f\left(T_{0}\right)\right)}{2},
\end{aligned}
$$

where $(\dagger)$ follows from Lemma $1,(\ddagger)$ from Lemma 2, $(\star)$ follows using the Chernoff-Hoeffding bound [69] and finally, (*) follows from the definition of $T_{0}$. Therefore, we have

$$
\begin{aligned}
& \mathbb{P}(\tau \leqslant l) \leqslant \frac{\bar{\Phi}\left(f\left(T_{0}\right)\right)}{2}+\Phi\left(f\left(T_{0}\right)\right)=1-\frac{\bar{\Phi}\left(f\left(T_{0}\right)\right)}{2} \\
\Longrightarrow & \mathbb{P}(\tau>l) \geqslant \frac{\bar{\Phi}\left(f\left(T_{0}\right)\right)}{2} .
\end{aligned}
$$

Taking the limit $l \rightarrow \infty$ and appealing to the continuity of probability, we obtain

$$
\begin{aligned}
& \mathbb{P}(\tau=\infty) \geqslant \frac{\bar{\Phi}\left(f\left(T_{0}\right)\right)}{2} \\
\Longrightarrow & \mathbb{P}\left(\bigcap_{m \geqslant 1} \bigcap_{a, b \in \mathcal{A}, a<b}\left\{\left|\mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)\right| \geqslant 4 \sqrt{m \log m}\right\}\right) \geqslant \frac{\bar{\Phi}\left(f\left(T_{0}\right)\right)}{2} .
\end{aligned}
$$

## B. 6 Proof of Theorem 10

We will initially assume $\delta>8 \sqrt{\log n / n}$ for technical convenience. In the final step leading up to the asserted bound, we will relax this assumption by offsetting regret appropriately.

Let $\mathcal{A}:=\{1, \ldots, K\}$. Define the following stopping times:

$$
\begin{aligned}
& \tau_{1}\left(s_{n}\right):=\inf \left\{m \geqslant s_{n}: \exists a, b \in \mathcal{A}, a<b \text { s.t. }\left|\mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)\right|<4 \sqrt{m \log m}\right\}, \\
& \tau_{2}\left(s_{n}\right):=\inf \left\{m \geqslant s_{n}:\left|\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)\right| \geqslant 4 \sqrt{m \log n} \forall a, b \in \mathcal{A}, a<b\right\} .
\end{aligned}
$$

To keep notations simple, we will suppress the argument and denote $\tau_{1}\left(s_{n}\right)$ and $\tau_{2}\left(s_{n}\right)$ by $\tau_{1}$ and $\tau_{2}$ respectively (the dependence on $s_{n}$ will be implicit going forward). Let $R_{t}$ denote the cumulative pseudo-regret of $\operatorname{ALG} 2(n)$ after $t \leqslant n$ pulls. Let D denote the event that the first batch of $K$ arms queried from the reservoir is "all-distinct," i.e., no two arms in this batch belong to the same type; let $D^{c}$ be the complement of this event. Let CI denote the event that the algorithm commits to an inferior-typed arm. Let $\tilde{R}$., $\bar{R}$. be independently drawn from the same distribution as $R$.. Let $x^{+}:=\max (x, 0)$ for $x \in \mathbb{R}$. Then, $R_{n}$ evolves according to the following stochastic
recursion:

$$
\begin{aligned}
& R_{n} \\
\leqslant & \mathbb{1}\{\mathrm{D}\}\left[\mathbb{1}\left\{\tau_{1}<\tau_{2}\right\}\left(\bar{\Delta} \min \left(K \tau_{1}, n\right)+\tilde{R}_{\left(n-K \tau_{1}\right)^{+}}\right)\right] \\
& \mathbb{1}\{\mathrm{D}\}\left[\mathbb{1}\left\{\tau_{1} \geqslant \tau_{2}\right\}\left(\bar{\Delta} \min \left(K \tau_{2}, n\right)+\mathbb{1}\{\mathrm{CI}\} \bar{\Delta}\left(n-K \tau_{2}\right)^{+}\right)\right] \\
& +\mathbb{1}\left\{\mathrm{D}^{\mathrm{C}}\right\}\left[\mathbb{1}\left\{\tau_{1}<\tau_{2}\right\}\left(\bar{\Delta} \min \left(K \tau_{1}, n\right)+\bar{R}_{\left(n-K \tau_{1}\right)^{+}}\right)+\mathbb{1}\left\{\tau_{1} \geqslant \tau_{2}\right\} \bar{\Delta} n\right] \\
\leqslant & \mathbb{1}\{\mathrm{D}\}\left[\mathbb{1}\left\{\tau_{1}<\tau_{2}\right\}\left(\bar{\Delta} \min \left(K \tau_{2}, n\right)+\tilde{R}_{n}\right)+\mathbb{1}\left\{\tau_{1} \geqslant \tau_{2}\right\}\left(\bar{\Delta} \min \left(K \tau_{2}, n\right)+\mathbb{1}\{\mathrm{CI}\} \bar{\Delta} n\right)\right] \\
& +\mathbb{1}\left\{\mathrm{D}^{\mathrm{C}}\right\}\left[\mathbb{1}\left\{\tau_{1}<\tau_{2}\right\}\left(\bar{\Delta} \min \left(K \tau_{1}, n\right)+\bar{R}_{n}\right)+\mathbb{1}\left\{\tau_{1} \geqslant \tau_{2}\right\} \bar{\Delta} n\right] \\
\leqslant & \mathbb{1}\{\mathrm{D}\}\left[2 \bar{\Delta} \min \left(K \tau_{2}, n\right)+\mathbb{1}\left\{\tau_{1}<\tau_{2}\right\} \tilde{R}_{n}+\mathbb{1}\{\mathrm{CI}\} \bar{\Delta} n\right]+\mathbb{1}\left\{\mathrm{D}^{\mathrm{C}}\right\}\left[\bar{\Delta} K \tau_{1}+\bar{R}_{n}+\mathbb{1}\left\{\tau_{1} \geqslant \tau_{2}\right\} \bar{\Delta} n\right] .
\end{aligned}
$$

Taking expectations on both sides, one recovers using the independence of $\tilde{R}_{n}, \bar{R}_{n}$ that

$$
\begin{aligned}
& \mathbb{E} R_{n} \\
\leqslant & \frac{\bar{\Delta}}{\mathbb{P}\left(\tau_{1} \geqslant \tau_{2} \mid \mathrm{D}\right)}\left[2 \mathbb{E}\left[\min \left(K \tau_{2}, n\right) \mid \mathrm{D}\right]+\left(\frac{\mathbb{P}\left(\mathrm{D}^{\mathrm{C}}\right)}{\mathbb{P}(\mathrm{D})}\right) \mathbb{E}\left[K \tau_{1} \mid \mathrm{D}^{\mathrm{C}}\right]\right] \\
& +\frac{\bar{\Delta}}{\mathbb{P}\left(\tau_{1} \geqslant \tau_{2} \mid \mathrm{D}\right)}\left[\left(\mathbb{P}(\mathrm{CI} \mid \mathrm{D})+\left(\frac{\mathbb{P}\left(\mathrm{D}^{\mathrm{C}}\right)}{\mathbb{P}(\mathrm{D})}\right) \mathbb{P}\left(\tau_{1} \geqslant \tau_{2} \mid \mathrm{D}^{\mathrm{C}}\right)\right) n\right],
\end{aligned}
$$

where $\mathbb{P}(\mathrm{D})=K!\prod_{i=1}^{K} \alpha_{i}$.

Lower bounding $\mathbb{P}\left(\tau_{1} \geqslant \tau_{2} \mid \mathrm{D}\right)$
Define the following:

$$
\begin{equation*}
\gamma\left(s_{n}\right):=\mathbb{P}\left(\tau_{1}\left(s_{n}\right)=\infty \mid \mathrm{D}\right), \tag{B.12}
\end{equation*}
$$

where $\tau_{1}\left(s_{n}\right)$ and D are as defined before. We will suppress the dependence on $s_{n}$ to keep notations minimal. Note that

$$
\begin{aligned}
\mathbb{P}\left(\tau_{1}<\tau_{2} \mid \mathrm{D}\right) & =\mathbb{P}\left(\tau_{1}<\tau_{2}, \tau_{2}=\infty \mid \mathrm{D}\right)+\mathbb{P}\left(\tau_{1}<\tau_{2}, \tau_{2}<\infty \mid \mathrm{D}\right) \\
& \leqslant \mathbb{P}\left(\tau_{2}=\infty \mid \mathrm{D}\right)+\mathbb{P}\left(\tau_{1}<\infty \mid \mathrm{D}\right) \\
& =\mathbb{P}\left(\tau_{1}<\infty \mid \mathrm{D}\right) \\
& =1-\gamma\left(s_{n}\right),
\end{aligned}
$$

where the equality in the third step follows since $\tau_{2}$ is almost surely finite on the event D (proved in §B. 6 below), and the final equality is due to (B.12). Thus, $\mathbb{P}\left(\tau_{1} \geqslant \tau_{2} \mid \mathrm{D}\right) \geqslant \gamma\left(s_{n}\right)$.

## Proof that $\tau_{2}$ is almost surely finite on D

Let $\mathbb{P}_{\mathrm{D}}(\cdot):=\mathbb{P}(\cdot \mid \mathrm{D})$ be the conditional measure w.r.t. the event D . Let $\mathcal{A}:=\{1, \ldots, K\}$. Then, by continuity of probability, we have

$$
\begin{aligned}
\mathbb{P}_{\mathrm{D}}\left(\tau_{2}=\infty\right) & =\lim _{l \rightarrow \infty} \mathbb{P}_{\mathrm{D}}\left(\tau_{2}>l\right) \\
& =\lim _{l \rightarrow \infty} \mathbb{P}_{\mathrm{D}}\left(\bigcap_{m=s_{n}}^{l} \bigcup_{a, b \in \mathcal{A}, a<b}\left\{\left|\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)\right|<4 \sqrt{m \log n}\right\}\right) \\
& \leqslant \lim _{l \rightarrow \infty} \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}_{\mathrm{D}}\left(\left|\sum_{j=1}^{l}\left(X_{a, j}-X_{b, j}\right)\right|<4 \sqrt{l \log n}\right) .
\end{aligned}
$$

On D, it must be that $\left|\mathbb{E}\left[X_{a, j}-X_{b, j}\right]\right| \geqslant \delta$. Without loss of generality, assume that $\mathbb{E}\left[X_{a, j}-X_{b, j}\right] \geqslant$
$\delta$. Then,

$$
\begin{aligned}
\mathbb{P}_{\mathrm{D}}\left(\tau_{2}=\infty\right) & \leqslant \lim _{l \rightarrow \infty} \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}_{\mathrm{D}}\left(\sum_{j=1}^{l}\left(X_{a, j}-X_{b, j}\right)<4 \sqrt{l \log n}\right) \\
& =\lim _{l \rightarrow \infty} \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}_{\mathrm{D}}\left(\sum_{j=1}^{l}\left(X_{a, j}-X_{b, j}-\delta\right)<-l\left(\delta-4 \sqrt{\frac{\log n}{l}}\right)\right) \\
& \leqslant \lim _{l \rightarrow \infty} \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}_{\mathrm{D}}\left(\sum_{j=1}^{l}\left(X_{a, j}-X_{b, j}-\delta\right)<-4 l \sqrt{\log n}\left(\frac{2}{\sqrt{n}}-\frac{1}{\sqrt{l}}\right)\right),
\end{aligned}
$$

where the last inequality follows since $\delta>8 \sqrt{\log n / n}$ (by assumption). Now, using the Chernoff-Hoeffding bound [69] together with the fact that $-1 \leqslant X_{a, j}-X_{b, j} \leqslant 1$, we obtain for $l>n$ and any $a, b \in \mathcal{A}, a<b$ that

$$
\begin{aligned}
\mathbb{P}_{\mathrm{D}}\left(\sum_{j=1}^{l}\left(X_{a, j}-X_{b, j}-\delta\right)<-4 l \sqrt{\log n}\left(\frac{2}{\sqrt{n}}-\frac{1}{\sqrt{l}}\right)\right) & \leqslant \exp \left[-8 l\left(\frac{2}{\sqrt{n}}-\frac{1}{\sqrt{l}}\right)^{2} \log n\right] \\
& =\exp \left[-8\left(\frac{4 l}{n}-4 \sqrt{\frac{l}{n}}+1\right)^{2} \log n\right] .
\end{aligned}
$$

Summing over $a, b \in \mathcal{A}, a<b$ and taking the limit $l \rightarrow \infty$ proves the stated assertion.

## Upper bounding $\mathbb{E}\left[\min \left(K \tau_{2}, n\right) \mid \mathrm{D}\right]$

Let $\mathbb{P}_{\mathrm{D}}(\cdot):=\mathbb{P}(\cdot \mid \mathrm{D})$ be the conditional measure w.r.t. the event D . Let $\mathcal{A}:=\{1, \ldots, K\}$. Then,

$$
\begin{aligned}
\mathbb{E}\left[\min \left(K \tau_{2}, n\right) \mid \mathrm{D}\right] & =K \mathbb{E}\left[\left.\min \left(\tau_{2}, \frac{n}{K}\right) \right\rvert\, \mathrm{D}\right] \\
& \leqslant K \mathbb{E}\left[\min \left(\tau_{2}, n\right) \mid \mathrm{D}\right] \\
& \leqslant K s_{n}+K \sum_{k=s_{n}+1}^{n} \mathbb{P}_{\mathrm{D}}\left(\tau_{2} \geqslant k\right) \\
& \leqslant K s_{n}+K \sum_{k=s_{n}}^{n} \mathbb{P}_{\mathrm{D}}\left(\tau_{2} \geqslant k+1\right) \\
& \leqslant K s_{n}+K \sum_{k=1}^{n} \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}_{\mathrm{D}}\left(\left|\sum_{j=1}^{k}\left(X_{a, j}-X_{b, j}\right)\right|<4 \sqrt{k \log n}\right) .
\end{aligned}
$$

On D, it must be that $\left|\mathbb{E}\left[X_{a, j}-X_{b, j}\right]\right| \geqslant \delta$. Without loss of generality, assume that $\mathbb{E}\left[X_{a, j}-X_{b, j}\right] \geqslant$ $\delta$. Then,

$$
\begin{aligned}
\mathbb{E}\left[\min \left(K \tau_{2}, n\right) \mid \mathrm{D}\right] & \leqslant K s_{n}+K \sum_{k=1}^{n} \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}_{\mathrm{D}}\left(\sum_{j=1}^{k}\left(X_{a, j}-X_{b, j}\right)<4 \sqrt{k \log n}\right) \\
& =K s_{n}+K \sum_{k=1}^{n} \sum_{a, b \in \mathcal{A}, a<b} \mathbb{P}_{\mathrm{D}}\left(\sum_{j=1}^{k}\left(X_{a, j}-X_{b, j}-\delta\right)<-k\left(\delta-4 \sqrt{\frac{\log n}{k}}\right)\right) \\
& \leqslant K s_{n}+\frac{32 K^{3} \log n}{\delta^{2}}+K \sum_{k=\left[\frac{64 \log n}{\delta^{2}}\right]^{a, b \in \mathcal{A}, a<b}} \mathbb{P}_{\mathrm{D}}\left(\sum_{j=1}^{k}\left(X_{a, j}-X_{b, j}-\delta\right)<\frac{-k \delta}{2}\right),
\end{aligned}
$$

where the last step follows since $\delta>8 \sqrt{\log n / n}$ (by assumption) implies $n>64 \log n / \delta^{2}$, and $k \geqslant 64 \log n / \delta^{2}$ implies $\delta-4 \sqrt{\log n / k} \geqslant \delta / 2$. Finally, using the Chernoff-Hoeffding inequality [69] together with the fact that $\left|X_{a, j}-X_{b, j}\right| \leqslant 1$, one obtains

$$
\mathbb{E}\left[\min \left(K \tau_{2}, n\right) \mid \mathrm{D}\right] \leqslant K s_{n}+\frac{32 K^{3} \log n}{\delta^{2}}+\frac{K^{3}}{2} \sum_{k=\left\lceil\frac{64 \log n}{\delta^{2}}\right\rceil}^{n} \exp \left(\frac{-\delta^{2} k}{8}\right) \leqslant K s_{n}+\frac{64 K^{3} \log n}{\delta^{2}}
$$

## Upper bounding $\mathbb{E}\left[K \tau_{1} \mid D^{c}\right]$

The event $D^{C}$ will be implicitly assumed and we will drop the conditional argument for notational simplicity. Let $\mathcal{A}:=\{1, \ldots, K\}$. Without loss of generality, suppose that arm 1 and 2 belong to the same type. Then,

$$
\begin{aligned}
\mathbb{E}\left[K \tau_{1} \mid \mathrm{D}^{\mathrm{c}}\right] & =K s_{n}+K \sum_{k \geqslant s_{n}+1} \mathbb{P}\left(\tau_{1} \geqslant k\right) \\
& =K s_{n}+K \sum_{k \geqslant s_{n}} \mathbb{P}\left(\tau_{1} \geqslant k+1\right) \\
& =K s_{n}+K \sum_{k \geqslant s_{n}} \mathbb{P}\left(\bigcap_{m=s_{n}}^{k} \bigcap_{a, b \in \mathcal{A}, a<b}\left\{\left|\mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{m \log m}\right\}\right) \\
& \leqslant K s_{n}+K \sum_{k \geqslant 1} \mathbb{P}\left(\left|\mathcal{Z}_{1,2}+\sum_{j=1}^{k}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{k \log k}\right) .
\end{aligned}
$$

Since $\mathcal{Z}_{a, b}$ is a standard Gaussian, and the increments $X_{1, j}-X_{2, j}$ are zero-mean sub-Gaussian with variance proxy 1 , it follows from the Chernoff-Hoeffding concentration bound [69] that

$$
\mathbb{E}\left[K \tau_{1} \mid D^{\mathrm{c}}\right] \leqslant K s_{n}+2 K \sum_{k \geqslant 1} \frac{1}{k^{4}}=\left(s_{n}+\frac{\pi^{4}}{45}\right) K<\left(s_{n}+3\right) K .
$$

## Upper bounding $\mathbb{P}(C I \mid D)$

Let $\mathbb{P}_{\mathrm{D}}(\cdot):=\mathbb{P}(\cdot \mid \mathrm{D})$ be the conditional measure w.r.t. the event D . Let $\mathcal{A}:=\{1, \ldots, K\}$ and without loss of generality, suppose that arm 1 is optimal (mean $\mu_{1}$ ). Then,

$$
\begin{aligned}
\mathbb{P}(\mathrm{CI} \mid \mathrm{D}) & \leqslant \mathbb{P}_{\mathrm{D}}\left(\bigcup_{b=2}^{K}\left\{\sum_{j=1}^{\tau_{2}}\left(X_{1, j}-X_{b, j}\right) \leqslant-4 \sqrt{\tau_{2} \log n}\right\}\right) \\
& \leqslant \sum_{b=2}^{K} \sum_{k=s_{n}}^{n} \mathbb{P}_{\mathrm{D}}\left(\sum_{j=1}^{k}\left(X_{1, j}-X_{b, j}\right) \leqslant-4 \sqrt{k \log n}\right)+\sum_{b=2}^{K} \mathbb{P}_{\mathrm{D}}\left(\tau_{2}>n\right) \\
& \leqslant \sum_{b=2}^{K} \sum_{k=1}^{n} \frac{1}{n^{8}}+\frac{K^{3}}{n^{8}} \\
& \leqslant \frac{K}{n^{7}}+\frac{K^{3}}{n^{8}},
\end{aligned}
$$

where the second-to-last step follows using the Chernoff-Hoeffding inequality [69].

## Upper bounding $\mathbb{P}\left(\tau_{1} \geqslant \tau_{2} \mid D^{c}\right)$

Let $\mathbb{P}_{D^{c}}(\cdot):=\mathbb{P}\left(\cdot \mid D^{c}\right)$ be the conditional measure w.r.t. the event $D^{c}$. Let $\mathcal{A}:=\{1, \ldots, K\}$. On $\mathrm{D}^{\mathrm{C}}$, there exist 2 arms in $\mathcal{A}$ that belong to the same type; without loss of generality suppose that these arms are indexed by 1,2 . Then,

$$
\begin{align*}
\mathbb{P}\left(\tau_{1} \geqslant \tau_{2} \mid D^{\mathrm{c}}\right) & \leqslant \mathbb{P}\left(\tau_{1}>n \mid \mathrm{D}^{\mathrm{c}}\right)+\mathbb{P}\left(\tau_{2} \leqslant n \mid D^{\mathrm{c}}\right) \\
& \leqslant \frac{2}{n^{4}}+\mathbb{P}_{\mathrm{D}^{\mathrm{c}}}\left(\tau_{2} \leqslant n\right) \\
& =\frac{2}{n^{4}}+\mathbb{P}_{\mathrm{D}^{\mathrm{c}}}\left(\bigcup_{m=s_{n}}^{n} \bigcap_{a, b \in \mathcal{A}, a<b}\left\{\left|\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)\right| \geqslant 4 \sqrt{m \log n}\right\}\right) \\
& \leqslant \frac{2}{n^{4}}+\sum_{m=1}^{n} \mathbb{P}_{\mathbb{D}^{\mathrm{c}}}\left(\left|\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{m \log n}\right) \\
& \leqslant \frac{2}{n^{4}}+\frac{2}{n^{7}}, \tag{B.13}
\end{align*}
$$

where the last step follows using the Chernoff-Hoeffding bound [69].

Putting everything together
Combining everything, one finally obtains that when $\delta>8 \sqrt{\log n / n}$,

$$
\mathbb{E} R_{n} \leqslant \frac{C K^{3} \bar{\Delta}}{\gamma\left(s_{n}\right)}\left(\frac{\log n}{\delta^{2}}+\frac{s_{n}}{\mathbb{P}(\mathrm{D})}\right),
$$

where $\gamma\left(s_{n}\right)$ is as defined in (B.12), $\mathbb{P}(\mathrm{D})=K!\prod_{i=1}^{K} \alpha_{i}$, and $C$ is some absolute constant. When $\delta \leqslant 8 \sqrt{\log n / n}$, regret is at most $\bar{\Delta} n \leqslant 64 \bar{\Delta} / \delta^{2} \log n$. Thus, the aforementioned bound, in fact, holds generally for some large enough absolute constant $C$.

## B. 7 Auxiliary results used in the analysis of ALG3

Lemma 3 (Persistence of heterogeneous consideration sets) Consider a two-armed bandit with rewards bounded in $[0,1]$, means $\mu_{1}>\mu_{2}$, and gap $\underline{\Delta}=\mu_{1}-\mu_{2}$. Let $\left\{X_{i, j}: j=1,2, \ldots\right\}$ denote the sequence of rewards collected from arm $i \in\{1,2\}$ by UCB1 [10]. Let $\mathcal{Z}$ be an independently generated standard Gaussian random variable. Let $\left(N_{1}(n), N_{2}(n)\right)$ be the per-arm sample counts under UCB1 up to and including time n. Define

$$
\begin{aligned}
& M_{n}:=\min \left(N_{1}(n), N_{2}(n)\right), \\
& \tau:=\inf \left\{n \in \mathbb{N}:\left|\mathcal{Z}+\sum_{j=1}^{M_{n}}\left(X_{1, j}-X_{2, j}\right)\right|<4 \sqrt{M_{n} \log M_{n}}\right\} .
\end{aligned}
$$

Then, $\mathbb{P}(\tau=\infty)>\beta_{\Delta, 2}$, where $\beta_{\Delta, 2}$ is as defined in (2.2) with $\delta \leftarrow \underline{\Delta}$ and $K \leftarrow 2$.

Lemma 4 (Fast rejection of homogeneous consideration sets) Consider a two-armed bandit where both arms have equal means. Let $\left\{X_{i, j}: j=1,2, \ldots\right\}$ denote the sequence of rewards collected from arm $i \in\{1,2\}$ by UCB1 [10]. Let $\mathcal{Z}$ be an independently generated standard Gaussian random variable. Let $\left(N_{1}(n), N_{2}(n)\right)$ be the per-arm sample counts under UCBI up to and including
time n. Define

$$
\begin{aligned}
& M_{n}:=\min \left(N_{1}(n), N_{2}(n)\right), \\
& \tau:=\inf \left\{n \in \mathbb{N}:\left|\mathcal{Z}+\sum_{j=1}^{M_{n}}\left(X_{1, j}-X_{2, j}\right)\right|<4 \sqrt{M_{n} \log M_{n}}\right\} .
\end{aligned}
$$

Then, there exists an absolute constant $C$ such that $\mathbb{E} \tau \leqslant C$.

## B.7.1 Proof of Lemma 3

Since the rewards are uniformly bounded in $[0,1]$, it follows that $N_{i}(n) \rightarrow \infty$ for each arm $i \in$ $\{1,2\}$ as $n \rightarrow \infty$ on every sample-path. This is due to the structure of the upper confidence bounds used by UCB1. Consequently, $M_{n}=\min \left(N_{1}(n), N_{2}(n)\right) \rightarrow \infty$ as $n \rightarrow \infty$ on every samplepath. Also note that $M_{n}$ is a weakly increasing integer-valued process (starting from 1) with unit increments, wherever they exist. Thus, it follows on every sample-path that $\tau$, in fact, weakly dominates the stopping time $\tau^{\prime}$ defined below

$$
\begin{equation*}
\tau^{\prime}:=\inf \left\{m \in \mathbb{N}:\left|\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right|<4 \sqrt{m \log m}\right\} \tag{B.14}
\end{equation*}
$$

Therefore, $\mathbb{P}(\tau=\infty) \geqslant \mathbb{P}\left(\tau^{\prime}=\infty\right)>\beta_{\underline{\Delta}, 2}$, where the last inequality follows from Proposition 1 with $\delta \leftarrow \underline{\Delta}$ and $K \leftarrow 2$.

## B.7.2 Proof of Lemma 4

Note that

$$
\begin{aligned}
\mathbb{E} \tau & =1+\sum_{k \geqslant 2} \mathbb{P}(\tau \geqslant k) \\
& =1+\sum_{k \geqslant 2} \mathbb{P}\left(\bigcap_{n=1}^{k-1}\left\{\left|\mathcal{Z}+\sum_{j=1}^{M_{n}}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{M_{n} \log M_{n}}\right\}\right) \\
& \leqslant 1+\sum_{k \geqslant 1} \mathbb{P}\left(\left|\mathcal{Z}+\sum_{j=1}^{M_{k}}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{M_{k} \log M_{k}}\right) \mid \\
& =1+\sum_{k \geqslant 1} \sum_{m=1}^{k} \mathbb{P}\left(\left|\mathcal{Z}+\sum_{j=1}^{M_{k}}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{M_{k} \log M_{k}}, N_{1}(k)=m\right) \\
& =1+\sum_{k \geqslant 1} \sum_{1 \leqslant m \leqslant k / 2} \mathbb{P}\left(\left|\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{m \log m}, N_{1}(k)=m\right) \\
& +\sum_{k \geqslant 1} \sum_{k / 2<m \leqslant k} \mathbb{P}\left(\left|\mathcal{Z}+\sum_{j=1}^{(k-m)}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{(k-m) \log (k-m)}, N_{1}(k)=m\right) \\
& =1+\sum_{k \geqslant 1} \sum_{1 \leqslant m \leqslant k / 2} \mathbb{P}\left(\left|\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{m \log m}, N_{1}(k)=m\right) \\
& +\sum_{k \geqslant 1} \sum_{1 \leqslant m<k / 2} \mathbb{P}\left(\left|\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{m \log m}, N_{2}(k)=m\right) \\
\leqslant & 1+2 \sum_{k \geqslant 1} \sum_{\theta k \leqslant m \leqslant k / 2} \mathbb{P}\left(\left|\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{m \log m}\right) \\
& +\sum_{k \geqslant 1}\left[\mathbb{P}\left(N_{1}(k) \leqslant \theta k\right)+\mathbb{P}\left(N_{2}(k) \leqslant \theta k\right)\right],
\end{aligned}
$$

where $\theta=1 / 2-\sqrt{15} / 8$. Using Theorem 4(i) of [21] with $\epsilon=\sqrt{15} / 8$, one obtains

$$
\begin{aligned}
\mathbb{E} \tau & \leqslant 1+2 \sum_{k \geqslant 1} \sum_{\theta k \leqslant m \leqslant k / 2} \mathbb{P}\left(\left|\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{m \log m}\right)+16 \sum_{k \geqslant 1} \frac{1}{k^{2}} \\
& \leqslant 1+4 \sum_{k \geqslant 1} \sum_{\theta k \leqslant m \leqslant k / 2} \frac{1}{m^{4}}+16 \sum_{k \geqslant 1} \frac{1}{k^{2}} \\
& \leqslant 1+\frac{4}{\theta^{4}} \sum_{k \geqslant 1} \frac{1}{k^{3}}+16 \sum_{k \geqslant 1} \frac{1}{k^{2}} .
\end{aligned}
$$

## B. 8 Proof of Theorem 11

Consider the first epoch and define the following:

$$
\begin{aligned}
& M_{n}:=\min \left(N_{1}(n), N_{2}(n)\right), \\
& \tau:=\inf \left\{n \in \mathbb{N}:\left|\mathcal{Z}+\sum_{j=1}^{M_{n}}\left(X_{1, j}-X_{2, j}\right)\right|<4 \sqrt{M_{n} \log M_{n}}\right\},
\end{aligned}
$$

where $\left(N_{1}(n), N_{2}(n)\right)$ are the per-arm sample counts under UCB1 up to and including time $n$. Note that $\tau$ marks the termination of epoch 1.

Let $R_{n}$ denote the cumulative pseudo-regret of ALG3 after $n$ pulls (superscript $\pi$ suppressed for notational convenience). Let $S_{n}$ denote the cumulative pseudo-regret of UCB1 after $n$ pulls in a two-armed bandit with gap $\underline{\Delta}$. Let $D$ and I respectively denote the events that the two arms queried in epoch 1 have distinct and identical types. Similarly, let OPT and INF respectively denote the events that the two arms have "optimal" and "inferior" types (Note that I = OPT U INF). Let $\tilde{R}_{n}, \bar{R}_{n}, \hat{R}_{n}$ be independently drawn from the same distribution as $R_{n}$. Then, note that $R_{n}$ admits the
following stochastic evolution:

$$
\begin{aligned}
R_{n} & =\mathbb{1}\{\mathrm{D}\}\left[S_{\min (\tau, n)}+\tilde{R}_{(n-\tau)^{+}}\right]+\mathbb{1}\{\mathrm{INF}\}\left[\underline{\Delta} \min (\tau, n)+\bar{R}_{(n-\tau)^{+}}\right]+\mathbb{1}\{\mathrm{OPT}\} \hat{R}_{(n-\tau)^{+}} \\
& \leqslant \mathbb{1}\{\mathrm{D}\}\left[S_{n}+\mathbb{1}\{\tau<n\} \tilde{R}_{n}\right]+\mathbb{1}\{\mathrm{INF}\}\left[\underline{\Delta} \tau+\bar{R}_{n}\right]+\mathbb{1}\{\mathrm{OPT}\} \hat{R}_{n} \\
& \leqslant \mathbb{1}\{\mathrm{D}\}\left[S_{n}+\mathbb{1}\{\tau<\infty\} \tilde{R}_{n}\right]+\mathbb{1}\{\mathrm{INF}\}\left[\underline{\Delta} \tau+\bar{R}_{n}\right]+\mathbb{1}\{\mathrm{OPT}\} \hat{R}_{n},
\end{aligned}
$$

where the first inequality follows since ALG3 is agnostic to $n$, and hence the pseudo-regret $R_{n}$ is weakly increasing in $n$. Taking expectations on both sides, one recovers using the independence of $\tilde{R}_{n}, \bar{R}_{n}, \hat{R}_{n}$ that

$$
\mathbb{E} R_{n} \leqslant \frac{1}{\mathbb{P}(\tau=\infty \mid \mathrm{D})}\left[\mathbb{E} S_{n}+\left(\frac{1-\alpha_{1}}{2 \alpha_{1}}\right) \underline{\Delta \mathbb{E}}[\tau \mid \mathrm{INF}]\right] \leqslant \frac{1}{\beta_{\underline{\Delta}, 2}}\left[\mathbb{E} S_{n}+C\left(\frac{1-\alpha_{1}}{2 \alpha_{1}}\right) \underline{\Delta}\right],
$$

where $\beta_{\Delta, 2}$ is as defined in (2.2) with $\delta \leftarrow \underline{\Delta}$ and $K \leftarrow 2$, and $C$ is some absolute constant; the last inequality follows using Lemma 3 and 4 . The stated assertion now follows since $\mathbb{E} S_{n} \leqslant$ $C^{\prime}(\log n / \underline{\Delta}+\underline{\Delta})$ for some absolute constant $C^{\prime}[10]$.

## B. 9 Auxiliary results used in the analysis of ALG4

Lemma 5 (Persistence of heterogeneous consideration sets) Consider a K-armed bandit with rewards bounded in $[0,1]$ and means $\mu_{1}>\ldots>\mu_{K}$. Let $\left\{X_{a, j}: j=1,2, \ldots\right\}$ denote the rewards collected from arm $a \in\{1, \ldots, K\}=: \mathcal{A}$ by UCB1 [10]. Let $\left\{\mathcal{Z}_{a, b}: a, b \in \mathcal{A}, a<b\right\}$ be a collection of $\binom{K}{2}$ independent standard Gaussian random variables. Let $N_{a}(n)$ be the sample count of arm a under UCB1 until time $n$. Define

$$
\begin{aligned}
M_{l} & :=\min _{a \in \mathcal{A}} N_{a}(l), \\
\tau & :=\inf \left\{l \geqslant K: \exists a, b \in \mathcal{A}, a<b \text { s.t. }\left|\mathcal{Z}_{a, b}+\sum_{j=1}^{M_{l}}\left(X_{a, j}-X_{b, j}\right)\right|<4 \sqrt{M_{l} \log M_{l}}\right\} .
\end{aligned}
$$

Then, $\mathbb{P}(\tau=\infty)>\beta_{\delta, K}$, where $\beta_{\delta, K}$ is as defined in (2.2).

Lemma 6 (Path-wise lower bound on the arm-sampling rate of UCB1) Consider a K-armed bandit with rewards bounded in $[0,1]$. Let $N_{a}(n)$ be the sample count of arm $a \in\{1, \ldots, K\}=: \mathcal{A}$ under UCB1 [10] until time $n$. Then, for all $n \geqslant K$,

$$
M_{n}:=\min _{a \in \mathcal{F}} N_{a}(n) \geqslant f(n),
$$

where $(f(n): n=K, K+1, \ldots)$ is some deterministic monotone non-decreasing integer-valued sequence satisfying $f(K)=1$ and $f(n) \rightarrow \infty$ as $n \rightarrow \infty$.

## B.9.1 Proof of Lemma 5

Suppose that there exists a sample-path on which some non-empty subset of arms $\mathfrak{H} \subset \mathcal{A}$ receives a bounded number of pulls asymptotically in the horizon of play. Also suppose that $\mathfrak{A}$ is the maximal such subset, i.e., each $\operatorname{arm}$ in $\mathcal{A} \backslash \mathfrak{A}$ is played infinitely often asymptotically on said sample-path. This implies that the UCB score of any $\operatorname{arm} \operatorname{in} \mathcal{A} \backslash \mathfrak{A}$ is at most $1+o(\sqrt{\log t})$ at time $t$ (since the empirical mean term remains bounded in $[0,1]$ ). At the same time, the boundedness hypothesis implies that the UCB score of any arm in $\mathfrak{A}$ is at least $\Omega(\sqrt{\log t})$. Thus, for $t$ large enough, UCB scores of arms in $\mathfrak{A}$ will start to dominate those in $\mathcal{A} \backslash \mathfrak{A}$ and the algorithm will end up playing an arm from $\mathfrak{A}$ at some point, thus increasing the cumulative sample-count of arms in $\mathfrak{A}$ by 1 . As $t$ grows further, one can replicate the preceding argument an arbitrary number of times to conclude that $\mathfrak{A}$ receives an unbounded number of pulls on the sample-path under consideration, thereby contradicting the boundedness hypothesis. Therefore, it must be the case that each arm in $\mathcal{A}$ is played infinitely often on every sample-path. Consequently, $M_{n}=\min _{a \in \mathcal{A}} N_{a}(n) \rightarrow \infty$ as $n \rightarrow \infty$ on every sample-path.

Now since $\left(M_{n}: n=K, K+1, \ldots\right)$ is an integer-valued process (starting from $M_{K}=1$ ) with unit increments (wherever they exist), it follows that on every sample-path, $\tau$, in fact, weakly
dominates the stopping time $\tau^{\prime}$ given by

$$
\begin{equation*}
\tau^{\prime}:=\inf \left\{m \in \mathbb{N}: \exists a, b \in \mathcal{A}, a<b \text { s.t. }\left|\mathcal{Z}_{a, b}+\sum_{j=1}^{m}\left(X_{a, j}-X_{b, j}\right)\right|<4 \sqrt{m \log m}\right\} . \tag{B.15}
\end{equation*}
$$

Therefore, $\mathbb{P}(\tau=\infty) \geqslant \mathbb{P}\left(\tau^{\prime}=\infty\right)>\beta_{\delta, K}$; the last inequality follows from Proposition 1 .

## B.9.2 Proof of Lemma 6

Suppose that $\mathcal{S}_{n}=\left\{\left(N_{a}(n): a \in \mathcal{A}\right)\right\}$ denotes the set of possible sample-count realizations under UCB1 when the horizon of play is $n$. Define $f(n):=\min _{\left(N_{a}(n): a \in \mathcal{A}\right) \in \mathcal{S}_{n}} \min _{a \in \mathcal{A}} N_{a}(n)$. Since $\mathcal{S}_{n}$ is finite, aforementioned minimum is attained at some $\left(N_{a}^{*}(n): a \in \mathcal{A}\right) \in \mathcal{S}_{n}$. Note that $\left(N_{a}^{*}(n): a \in \mathcal{A}\right)$ is not a random vector as it corresponds to a specific set of sample-paths (possibly non-unique) on which $\min _{a \in \mathcal{A}} N_{a}(n)$ is minimized. Therefore, $f(n)=\min _{a \in \mathcal{A}} N_{a}^{*}(n)$ is deterministic. We have already established in the proof of Lemma 5 that for each $a \in \mathcal{A}$, $N_{a}(n) \rightarrow \infty$ as $n \rightarrow \infty$ on every sample-path. In particular, this also implies $N_{a}^{*}(n) \rightarrow \infty$ as $n \rightarrow \infty$. Thus, we have established the existence of a sequence $f(n)$ satisfying the assertions of the lemma.

## B. 10 Proof of Theorem 12

Let $\mathcal{A}:=\{1, \ldots, K\}$ be the collection of $K$ arms queried during the first epoch. Consider an arbitrary $l \in \mathbb{N}$ s.t. $l \geqslant K$ and define the following:

$$
\begin{aligned}
& M_{l}:=\min _{a \in \mathcal{A}} N_{a}(l), \\
& \tau:=\inf \left\{l \geqslant K: \exists a, b \in \mathcal{A}, a<b \text { s.t. }\left|\mathcal{Z}_{a, b}+\sum_{j=1}^{M_{l}}\left(X_{a, j}-X_{b, j}\right)\right|<4 \sqrt{M_{l} \log M_{l}}\right\},
\end{aligned}
$$

where $N_{a}(l)$ denotes the sample count of arm $a$ under UCB1 until time $l$, and $\tau$ marks the termination of epoch 1 . Let $R_{n}$ denote the cumulative pseudo-regret of ALG4 after $n$ pulls (superscript
$\pi$ suppressed for notational convenience). Let $S_{n}$ denote the cumulative pseudo-regret of UCB1 after $n$ pulls in a $K$-armed bandit with means $\mu_{1}>\mu_{2}>\ldots>\mu_{K}$. Let $D$ denote the event that the $K$ arms queried in epoch 1 have distinct types (no two belong to the same type). Let $\tilde{R}_{n}, \bar{R}_{n}$ be independently drawn from the same distribution as $R_{n}$. Then, the evolution of $R_{n}$ satisfies

$$
\begin{aligned}
R_{n} & \leqslant \mathbb{1}\{\mathrm{D}\}\left[S_{\min (\tau, n)}+\tilde{R}_{(n-\tau)^{+}}\right]+\mathbb{1}\left\{\mathrm{D}^{\mathrm{C}}\right\}\left[\bar{\Delta} \min (\tau, n)+\bar{R}_{(n-\tau)^{+}}\right] \\
& \leqslant \mathbb{1}\{\mathrm{D}\}\left[S_{n}+\mathbb{1}\{\tau<n\} \tilde{R}_{n}\right]+\mathbb{1}\left\{\mathrm{D}^{\mathrm{c}}\right\}\left[\bar{\Delta} \min (\tau, n)+\bar{R}_{n}\right] \\
& \leqslant \mathbb{1}\{\mathrm{D}\}\left[S_{n}+\mathbb{1}\{\tau<\infty\} \tilde{R}_{n}\right]+\mathbb{1}\left\{\mathrm{D}^{\mathrm{c}}\right\}\left[\bar{\Delta} \min (\tau, n)+\bar{R}_{n}\right],
\end{aligned}
$$

where $(\dagger)$ follows since ALG 4 is agnostic to $n$, and hence the pseudo-regret $R_{n}$ is weakly increasing in $n$. Taking expectations on both sides, one recovers using the independence of $\tilde{R}_{n}, \bar{R}_{n}$ that

$$
\mathbb{E} R_{n} \leqslant \frac{1}{\mathbb{P}(\tau=\infty \mid \mathrm{D})}\left[\mathbb{E} S_{n}+\left(\frac{\bar{\Delta} \mathbb{E}\left[\min (\tau, n) \mid \mathrm{D}^{\mathrm{C}}\right]}{\mathbb{P}(\mathrm{D})}\right)\right] \leqslant \frac{1}{\beta_{\delta, K}}\left[\mathbb{E} S_{n}+\left(\frac{\bar{\Delta} \mathbb{E}\left[\min (\tau, n) \mid \mathrm{D}^{\mathrm{C}}\right]}{\mathbb{P}(\mathrm{D})}\right)\right],
$$

where $\mathbb{P}(\mathrm{D})=K!\prod_{i=1}^{K} \alpha_{i}$, and the last inequality follows using Lemma 5 with $\beta_{\delta, K}$ as defined in (2.2). We know that $\mathbb{E} S_{n} \leqslant C K(\log n / \underline{\Delta}+\bar{\Delta})$ for some absolute constant $C$ [10]. The rest of the proof is geared towards showing that $\mathbb{E}\left[\min (\tau, n) \mid \mathrm{D}^{\mathrm{C}}\right]=o(n)$.

Proof of $\mathbb{E}\left[\min (\tau, n) \mid D^{\mathrm{c}}\right]=o(n)$

Let $\mathbb{P}_{D^{c}}(\cdot):=\mathbb{P}\left(\cdot \mid D^{\mathfrak{c}}\right)$ be the conditional measure w.r.t. $D^{c}$. On $D^{c}$, there exist two arms in the consideration set $\mathcal{A}$ that belong to the same type. Without loss of generality, suppose these are
indexed 1 and 2. Then,

$$
\begin{aligned}
\mathbb{E}\left[\min (\tau, n) \mid D^{\mathrm{c}}\right] & \leqslant K+\sum_{k=K+1}^{n} \mathbb{P}_{\mathrm{D}}(\tau \geqslant k) \\
& \leqslant K+\sum_{k=K}^{n} \mathbb{P}_{\mathrm{D}} \mathrm{c}(\tau \geqslant k+1) \\
& =K+\sum_{k=K}^{n} \mathbb{P}_{D^{\mathrm{c}}}\left(\bigcap_{l=1}^{k} \bigcap_{a, b \in \mathcal{A}, a<b}\left\{\left|\mathcal{Z}_{a, b}+\sum_{j=1}^{M_{l}}\left(X_{a, j}-X_{b, j}\right)\right| \geqslant 4 \sqrt{M_{l} \log M_{l}}\right\}\right) \\
& \leqslant K+\sum_{k=K}^{n} \mathbb{P}_{\mathrm{D}}\left(\left|\mathcal{Z}_{1,2}+\sum_{j=1}^{M_{k}}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{M_{k} \log M_{k}}\right) \\
& =K+\sum_{k=K}^{n} \sum_{m=1}^{k} \mathbb{P}_{\mathrm{D}}\left(\left|\mathcal{Z}_{1,2}+\sum_{j=1}^{M_{k}}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{M_{k} \log M_{k}}, M_{k}=m\right) \\
& =K+\sum_{k=K}^{n} \sum_{m=f(k)}^{k} \mathbb{P}_{D^{\mathrm{c}}}\left(\left|\mathcal{Z}_{1,2}+\sum_{j=1}^{M_{k}}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{M_{k} \log M_{k}}, M_{k}=m\right) \\
& \leqslant K+\sum_{k=K}^{n} \sum_{m=f(k)}^{k} \mathbb{P}_{\mathrm{D}}\left(\left|\mathcal{Z}_{1,2}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{m \log m}\right) \\
& \leqslant K+2 \sum_{(\ddagger=K}^{n} \sum_{m=f(k)}^{k} \frac{1}{m^{4}} \\
& =K+2 \sum_{k=K}^{n}\left(\frac{1}{(f(k))^{4}}+\sum_{m=f(k)+1}^{k} \frac{1}{m^{4}}\right) \\
& \leqslant K+2 \sum_{k=K}^{n}\left(\frac{1}{(f(k))^{4}}+\frac{1}{3(f(k))^{3}}\right),
\end{aligned}
$$

where $(\dagger)$ follows from Lemma 6 , and $(\ddagger)$ using the Chernoff-Hoeffding bound [69]. Since $f(k)$ is monotone non-decreasing and coercive in $k$, it follows that $\mathbb{E}\left[\min (\tau, n) \mid D^{\mathrm{c}}\right]=o(n)$, where the little-Oh only hides dependence on $K$.

## Appendix C: Appendix to Chapter 4

## General organization

1. Appendix C. 1 provides the proof of Theorem 14.
2. Appendix C. 2 provides the proof of Proposition 2.
3. Appendix C. 3 provides the regret analysis framework for Algorithm 4.
4. Appendix C.3.1 provides the proof of Theorem 15.
5. Appendix C.3.2 provides the proof of Theorem 17.
6. Appendix C. 4 states the auxiliary results used in regret analysis of Algorithm 5.
7. Appendix C. 5 provides the regret analysis framework for Algorithm 5.
8. Appendix C.5.1 provides the proof of Theorem 16.
9. Appendix C.5.2 provides the proof of Theorem 18.

## C. 1 Proof of Theorem 14

## C.1.1 Proof for Model 1

In order to prove this result, we consider an oracle that can perfectly observe whether an arm is "optimal" or "inferior"-typed immediately upon pulling it. If such an oracle incurs linear regret, then every policy that only gets to observe a noisy realization of the mean rewards associated with the types, must necessarily incur linear regret as well.

Clearly, the optimal oracle policy $\pi^{*}$ is one that keeps pulling new arms until it finds one of the optimal type (type 1), which it then persists with for the remaining duration of play. Let $Y$ denote the time at which an arm of the optimal type is pulled for the first time under $\pi^{*}$. Then,

$$
\begin{equation*}
\mathbb{P}(Y \geqslant k)=\prod_{t=1}^{k-1}(1-\alpha(t)) \quad \text { for } k \geqslant 2, \quad \mathbb{P}(Y \geqslant 1)=1 . \tag{C.1}
\end{equation*}
$$

The expected cumulative regret of the aforementioned policy at time $n$ is

$$
\mathbb{E} R_{n}^{\pi^{*}}=\sum_{k=1}^{n} \mathbb{P}(Y=k) \Delta(k-1)+\mathbb{P}(Y>n) \Delta n>\mathbb{P}(Y>n) \Delta n>\mathbb{P}(Y=\infty) \Delta n
$$

Thus, if $\mathbb{P}(Y=\infty)$ is bounded away from 0 , linear regret is unavoidable. Since $\lim _{t \rightarrow \infty} \alpha(t)=$ 0 , we know that $\exists t_{0} \in \mathbb{N}$ s.t. $\alpha(t)<1 / 2$ for all $t>t_{0}$. Then,

$$
\begin{align*}
\mathbb{P}(Y=\infty)=\prod_{t=1}^{\infty}(1-\alpha(t))=\exp \left(\sum_{t=1}^{\infty} \log (1-\alpha(t))\right) & =\prod_{t=1}^{t_{0}}(1-\alpha(t)) \exp \left(\sum_{t=t_{0}+1}^{\infty} \log (1-\alpha(t))\right) \\
& >\prod_{t=1}^{t_{0}}(1-\alpha(t)) \exp \left(-2 \sum_{t=t_{0}+1}^{\infty} \alpha(t)\right), \tag{C.2}
\end{align*}
$$

where the final inequality follows since $\alpha(t)<1 / 2$ for $t>t_{0}$ and $\log (1-x)>-2 x$ for $x \in(0,1 / 2]$. Since $t_{0}$ is finite, it is clear from (C.2) that a sufficient condition for $\mathbb{P}(Y=\infty)$ to be bounded away from 0 is the summability of $\alpha(t)$, i.e., $\sum_{t \in \mathbb{N}} \alpha(t)<\infty$.

## C.1.2 Proof for Model 2

The proof for this model proceeds along similar lines as the above. One starts by considering an oracle that observes the type of a queried arm perfectly and immediately. The structure of the optimal oracle policy can be argued to be identical to the one discussed in the previous section. Subsequent steps of the proof are instructive.

## C. 2 Proof of Proposition 2

Consider the following stopping time:

$$
\tau:=\inf \left\{m \in \mathbb{N}: \mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)<4 \sqrt{m \log m}\right\} .
$$

Since $\mathbb{P}\left(\bigcap_{m \geqslant 1}\left\{\left|\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{m \log m}\right\}\right) \geqslant \mathbb{P}(\tau=\infty)$, it suffices to show that $\mathbb{P}(\tau=\infty)$ is bounded away from 0 . To this end, define the following:

$$
\begin{aligned}
T_{0} & :=\left[\left(\frac{64}{\Delta^{2}}\right) \log ^{2}\left(\frac{64}{\Delta^{2}}\right)\right], \\
f(x) & :=x+4 \sqrt{x \log x} \quad \text { for } x \geqslant 1 .
\end{aligned}
$$

Lemma 7 It is the case that

$$
\left\{\mathcal{Z}>f\left(T_{0}\right)\right\} \subseteq \bigcap_{m=1}^{T_{0}}\left\{\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right) \geqslant 4 \sqrt{m \log m}\right\}
$$

Proof of Lemma 7. Note that

$$
\begin{aligned}
\mathcal{Z} & >f\left(T_{0}\right) \\
& =T_{0}+4 \sqrt{T_{0} \log T_{0}} \\
& \geqslant m+4 \sqrt{m \log m} \forall 1 \leqslant m \leqslant T_{0} \\
& \geqslant \sum_{j=1}^{m}\left(X_{2, j}-X_{1, j}\right)+4 \sqrt{m \log m} \forall 1 \leqslant m \leqslant T_{0} \\
\Longrightarrow \mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right) & \geqslant 4 \sqrt{m \log m} \forall 1 \leqslant m \leqslant T_{0},
\end{aligned}
$$

where $(\mathfrak{a})$ follows since the rewards are bounded in $[0,1]$, i.e., $\left|X_{1, j}-X_{2, j}\right| \leqslant 1$.

Lemma 8 For $m \geqslant T_{0}$, it is the case that

$$
\Delta \geqslant 8 \sqrt{\frac{\log m}{m}}
$$

Proof of Lemma 8. First of all, note that $T_{0} \geqslant 64$ (since $\Delta \leqslant 1$ ). For $s=\left(64 / \Delta^{2}\right) \log ^{2}\left(64 / \Delta^{2}\right)$, one has

$$
\Delta^{2}=\frac{64 \log ^{2}\left(\frac{64}{\Delta^{2}}\right)}{s} \underset{(\mathfrak{b})}{\geqslant} \frac{64\left[\log \left(\frac{64}{\Delta^{2}}\right)+2 \log \log \left(\frac{64}{\Delta^{2}}\right)\right]}{s}=\frac{64 \log s}{s}
$$

where (b) follows since the function $g(x):=x^{2}-x-2 \log x$ is monotone increasing for $x \geqslant \log 64$ (think of $\log \left(64 / \Delta^{2}\right)$ as $x$ ), and therefore attains its minimum at $x=\log 64$; one can verify that this minimum is strictly positive. Furthermore, since $\log s / s$ is monotone decreasing for $s \geqslant 64$, it follows that for any $m \geqslant T_{0}$,

$$
\Delta^{2} \geqslant \frac{64 \log m}{m}
$$

Now coming back to the proof of Proposition 2, consider an arbitrary $l \in \mathbb{N}$ such that $l>T_{0}$. Then,

$$
\begin{aligned}
\mathbb{P}(\tau \leqslant l) & =\mathbb{P}\left(\tau \leqslant l, \mathcal{Z}>f\left(T_{0}\right)\right)+\mathbb{P}\left(\tau \leqslant l, \mathcal{Z} \leqslant f\left(T_{0}\right)\right) \\
& \leqslant \mathbb{P}\left(\tau \leqslant l, \mathcal{Z}>f\left(T_{0}\right)\right)+\Phi\left(f\left(T_{0}\right)\right)
\end{aligned}
$$

Now,

$$
\begin{aligned}
& \mathbb{P}\left(\tau \leqslant l, \mathcal{Z}>f\left(T_{0}\right)\right)=\mathbb{P}\left(\bigcup_{m=1}^{l}\left\{\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)<4 \sqrt{m \log m}, \mathcal{Z}>f\left(T_{0}\right)\right\}\right) \\
& \underset{(\dagger)}{=} \mathbb{P}\left(\bigcup_{m=T_{0}}^{l}\left\{\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)<4 \sqrt{m \log m}, \mathcal{Z}>f\left(T_{0}\right)\right\}\right) \\
& \leqslant \sum_{m=T_{0}}^{l} \mathbb{P}\left(\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)<4 \sqrt{m \log m}, \mathcal{Z}>f\left(T_{0}\right)\right) \\
& =\sum_{m=T_{0}}^{l} \mathbb{P}\left(\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}-\Delta\right)<-m\left(\Delta-4 \sqrt{\frac{\log m}{m}}\right), \mathcal{Z}>f\left(T_{0}\right)\right) \\
& \leqslant \sum_{m=T_{0}}^{l} \mathbb{P}\left(\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}-\Delta\right)<-m\left(\Delta-4 \sqrt{\frac{\log m}{m}}\right), \mathcal{Z}>f\left(T_{0}\right)\right) \\
& \underset{\text { ( } \ddagger)}{\leqslant} \sum_{m=T_{0}}^{l} \mathbb{P}\left(\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}-\Delta\right)<-4 \sqrt{m \log m}, \mathcal{Z}>f\left(T_{0}\right)\right) \\
& =\bar{\Phi}\left(f\left(T_{0}\right)\right) \sum_{m=T_{0}}^{l} \mathbb{P}\left(\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}-\Delta\right)<-4 \sqrt{m \log m}\right) \\
& \underset{(\star)}{\leqslant} \bar{\Phi}\left(f\left(T_{0}\right)\right) \sum_{m=T_{0}}^{l} \frac{1}{m^{8}} \\
& \underset{(*)}{\leqslant} \bar{\Phi}\left(f\left(T_{0}\right)\right) \sum_{m=2}^{\infty} \frac{1}{m^{8}} \\
& \underset{(\bullet)}{=} \bar{\Phi}\left(f\left(T_{0}\right)\right)(\zeta(8)-1) \\
& \leqslant \frac{\bar{\Phi}\left(f\left(T_{0}\right)\right)}{200},
\end{aligned}
$$

where $(\dagger)$ follows from Lemma 7, $(\ddagger)$ from Lemma $8,(\star)$ follows using the Chernoff-Hoeffding bound [69], $(*)$ since $T_{0}>1$ (this is because $\Delta \leqslant 1$ ) and finally in $(\bullet), \zeta(\cdot)$ represents the Riemann
zeta function. Therefore, we have

$$
\begin{aligned}
& \mathbb{P}(\tau \leqslant l) \leqslant \frac{\bar{\Phi}\left(f\left(T_{0}\right)\right)}{200}+\Phi\left(f\left(T_{0}\right)\right)=1-\left(\frac{199}{200}\right) \bar{\Phi}\left(f\left(T_{0}\right)\right) \\
\Longrightarrow & \mathbb{P}(\tau>l) \geqslant\left(\frac{199}{200}\right) \bar{\Phi}\left(f\left(T_{0}\right)\right) .
\end{aligned}
$$

Taking the limit $l \rightarrow \infty$ and appealing to the continuity of probability, we obtain

$$
\begin{aligned}
& \mathbb{P}(\tau=\infty) \geqslant\left(\frac{199}{200}\right) \bar{\Phi}\left(f\left(T_{0}\right)\right) \\
\Longrightarrow & \mathbb{P}\left(\bigcap_{m \geqslant 1}\left\{\left|\mathcal{Z}+\sum_{j=1}^{m}\left(X_{1, j}-X_{2, j}\right)\right| \geqslant 4 \sqrt{m \log m}\right\}\right) \geqslant\left(\frac{199}{200}\right) \bar{\Phi}\left(f\left(T_{0}\right)\right) .
\end{aligned}
$$

## C. 3 Analysis of Algorithm 4

The horizon of play is divided into epochs of length $2 m=2\left\lceil\left(2 / \delta^{2}\right) \log n\right\rceil$ each (exactly one pair of arms is played $m$ times each per epoch), e.g., epoch 1 starts at $t=1$, epoch 2 at $t=$ $2 m+1$, and so on. The decision to commit forever to an empirically superior arm or to discard the consideration set of arms and reinitialize the policy, is taken after an epoch ends. For each $k \geqslant 1$, let $S_{k}$ denote the cumulative pseudo-regret incurred by Algorithm 1 when it is initialized at the beginning of epoch $k$ and run until the end of the horizon, i.e., from $t=(2 k-2) m+1$ to $t=n$. Let $S_{k}^{\prime}$ denote an i.i.d. copy of $S_{k}$. We are interested in an upper bound on $\mathbb{E} R_{n}^{\pi}=\mathbb{E} S_{1}$, where $\pi=$ Algorithm 1. To this end, suppose that $\pi$ is initialized at time $(2 k-2) m+1$ (beginning of epoch $k$ ). Label the arms played in this epoch as $\{1,2\}$ (arm 1 is played first). For $i \in\{1,2\}$, let $\bar{X}_{i}$ denote the empirical mean reward from the $m$ plays of arm $i$ in this epoch. Let $\kappa(i) \in \mathcal{K}=\{1,2\}$ denote the type of arm $i$. Recall that type 1 is optimal and that, the probability of a new arm queried from the reservoir at time $t$ being optimal-typed is $\alpha(t)$. Let $\mathbb{1}\{E\}$ denote the indicator corresponding to
an event $E$. Consider the following events:

$$
\begin{align*}
A & :=\{\kappa(1)=1, \kappa(2)=2\},  \tag{C.3}\\
B & :=\{\kappa(1)=2, \kappa(2)=1\},  \tag{C.4}\\
C & :=\{\kappa(1)=2, \kappa(2)=2\},  \tag{C.5}\\
D & :=\{\kappa(1)=1, \kappa(2)=1\}, \tag{C.6}
\end{align*}
$$

where $\kappa(1)$ and $\kappa(2)$ are independent random variables with distributions given by $\mathbb{P}(\kappa(1)=1)=$ $\alpha((2 k-2) m+1)=: \alpha_{k}$ and $\mathbb{P}(\kappa(2)=1)=\alpha((2 k-2) m+2)=: \tilde{\alpha}_{k}$ respectively. Now, observe that $S_{k}$ evolves according to the following stochastic recursion:

$$
\begin{align*}
S_{k}= & \mathbb{1}\{A\}\left[\Delta m+\mathbb{1}\left\{\bar{X}_{2}-\bar{X}_{1}>\delta\right\} \Delta(n-2 k m)+\left\{\left|\bar{X}_{1}-\bar{X}_{2}\right|<\delta\right\} S_{k+1}^{\prime}\right] \\
& +\mathbb{1}\{B\}\left[\Delta m+\mathbb{1}\left\{\bar{X}_{1}-\bar{X}_{2}>\delta\right\} \Delta(n-2 k m)+\mathbb{1}\left\{\left|\bar{X}_{1}-\bar{X}_{2}\right|<\delta\right\} S_{k+1}^{\prime}\right] \\
& +\mathbb{1}\{C\}\left[2 \Delta m+\mathbb{1}\left\{\left|\bar{X}_{1}-\bar{X}_{2}\right|>\delta\right\} \Delta(n-2 k m)+\mathbb{1}\left\{\left|\bar{X}_{1}-\bar{X}_{2}\right|<\delta\right\} S_{k+1}^{\prime}\right] \\
& +\mathbb{1}\{D\} \mathbb{1}\left\{\left|\bar{X}_{1}-\bar{X}_{2}\right|<\delta\right\} S_{k+1}^{\prime} . \tag{C.7}
\end{align*}
$$

Collecting like terms in (C.7) together,

$$
\begin{align*}
S_{k}= & \mathbb{1}\{A\} \mathbb{1}\left\{\bar{X}_{2}-\bar{X}_{1}>\delta\right\} \Delta(n-2 k m)+\mathbb{1}\{B\} \mathbb{1}\left\{\bar{X}_{1}-\bar{X}_{2}>\delta\right\} \Delta(n-2 k m) \\
& +\mathbb{1}\{C\} \mathbb{1}\left\{\left|\bar{X}_{1}-\bar{X}_{2}\right|>\delta\right\} \Delta(n-2 k m)+[\mathbb{1}\{A \cup B\}+2 \mathbb{1}\{C\}] \Delta m+\mathbb{1}\left\{\left|\bar{X}_{1}-\bar{X}_{2}\right|<\delta\right\} S_{k+1}^{\prime} . \tag{C.8}
\end{align*}
$$

Define the following conditional events:

$$
\begin{align*}
& E_{1}:=\left\{\bar{X}_{2}-\bar{X}_{1}>\delta \mid A\right\},  \tag{C.9}\\
& E_{2}:=\left\{\bar{X}_{1}-\bar{X}_{2}>\delta \mid B\right\},  \tag{C.10}\\
& E_{3}:=\left\{\left|\bar{X}_{1}-\bar{X}_{2}\right|>\delta \mid C\right\},  \tag{C.11}\\
& E_{4}:=\left\{\left|\bar{X}_{1}-\bar{X}_{2}\right|<\delta \mid C \cup D\right\},  \tag{C.12}\\
& E_{5}:=\left\{\left|\bar{X}_{1}-\bar{X}_{2}\right|<\delta \mid A \cup B\right\} . \tag{C.13}
\end{align*}
$$

Taking expectations on both sides in (C.8) and rearranging, one obtains the following using (C.9)(C.13):

$$
\begin{align*}
& \mathbb{E} S_{k} \\
&= {\left[\alpha_{k}\left(1-\tilde{\alpha}_{k}\right) \mathbb{P}\left(E_{1}\right)+\tilde{\alpha}_{k}\left(1-\alpha_{k}\right) \mathbb{P}\left(E_{2}\right)\right] \Delta(n-2 k m)+\left[\left(1-\alpha_{k}\right)\left(1-\tilde{\alpha}_{k}\right) \mathbb{P}\left(E_{3}\right)\right] \Delta(n-2 k m) } \\
&+\left[\alpha_{k}\left(1-\tilde{\alpha}_{k}\right)+\tilde{\alpha}_{k}\left(1-\alpha_{k}\right)+2\left(1-\alpha_{k}\right)\left(1-\tilde{\alpha}_{k}\right)\right] \Delta m+\mathbb{P}\left(\left|\bar{X}_{1}-\bar{X}_{2}\right|<\delta\right) \mathbb{E} S_{k+1} . \tag{C.14}
\end{align*}
$$

Note that (C.14) follows from (C.8) since $\mathbb{E}\left[\mathbb{1}\left\{\left|\bar{X}_{1}-\bar{X}_{2}\right|<\delta\right\} S_{k+1}^{\prime}\right]=\mathbb{P}\left(\left|\bar{X}_{1}-\bar{X}_{2}\right|<\delta\right) \mathbb{E} S_{k+1}$ due to the independence of $S_{k+1}^{\prime}$. Further note that

$$
\begin{equation*}
\mathbb{P}\left(\left|\bar{X}_{1}-\bar{X}_{2}\right|<\delta\right)=\left[\alpha_{k} \tilde{\alpha}_{k}+\left(1-\alpha_{k}\right)\left(1-\tilde{\alpha}_{k}\right)\right] \mathbb{P}\left(E_{4}\right)+\left[\alpha_{k}\left(1-\tilde{\alpha}_{k}\right)+\tilde{\alpha}_{k}\left(1-\alpha_{k}\right)\right] \mathbb{P}\left(E_{5}\right) \tag{C.15}
\end{equation*}
$$

From (C.14) and (C.15), it follows after a little rearrangement that

$$
\begin{equation*}
\mathbb{E} S_{k}=\xi_{1}(k)-k \xi_{2}(k)+\xi_{3}(k) \mathbb{E} S_{k+1}, \tag{C.16}
\end{equation*}
$$

where the $\xi_{i}(k)$ 's are given by

$$
\begin{align*}
\xi_{1}(k):= & \Delta\left[\alpha_{k}\left(1-\tilde{\alpha}_{k}\right) \mathbb{P}\left(E_{1}\right)+\tilde{\alpha}_{k}\left(1-\alpha_{k}\right) \mathbb{P}\left(E_{2}\right)\right] n+\Delta\left[\left(1-\alpha_{k}\right)\left(1-\tilde{\alpha}_{k}\right) \mathbb{P}\left(E_{3}\right)\right] n \\
& +\Delta\left(2-\alpha_{k}-\tilde{\alpha}_{k}\right) m,  \tag{C.17}\\
\xi_{2}(k):= & 2 \Delta\left[\alpha_{k}\left(1-\tilde{\alpha}_{k}\right) \mathbb{P}\left(E_{1}\right)+\tilde{\alpha}_{k}\left(1-\alpha_{k}\right) \mathbb{P}\left(E_{2}\right)\right] m+2 \Delta\left[\left(1-\alpha_{k}\right)\left(1-\tilde{\alpha}_{k}\right) \mathbb{P}\left(E_{3}\right)\right] m,  \tag{C.18}\\
\xi_{3}(k):= & {\left[\alpha_{k} \tilde{\alpha}_{k}+\left(1-\alpha_{k}\right)\left(1-\tilde{\alpha}_{k}\right)\right] \mathbb{P}\left(E_{4}\right)+\left[\alpha_{k}\left(1-\tilde{\alpha}_{k}\right)+\tilde{\alpha}_{k}\left(1-\alpha_{k}\right)\right] \mathbb{P}\left(E_{5}\right) . } \tag{C.19}
\end{align*}
$$

Observe that the recursion in (C.16) is solvable in closed-form and admits the following solution:

$$
\begin{equation*}
\mathbb{E} S_{1}=\sum_{k=1}^{l}\left(\xi_{1}(k) \prod_{j=1}^{k-1} \xi_{3}(j)\right)-\sum_{k=1}^{l}\left(k \xi_{2}(k) \prod_{j=1}^{k-1} \xi_{3}(j)\right)+\mathbb{E} S_{l+1}\left(\prod_{k=1}^{l} \xi_{3}(k)\right), \tag{C.20}
\end{equation*}
$$

where $l:=\lfloor n /(2 m)\rfloor,\lfloor\cdot\rfloor$ being the "floor" operator. Since the $\xi_{i}(k)$ 's are non-negative for all $i \in\{1,2,3\}, k \in \mathbb{N}$ and $\mathbb{E} S_{l+1} \leqslant 2 \Delta m$, it follows that

$$
\begin{equation*}
\mathbb{E} R_{n}^{\pi}=\mathbb{E} S_{1} \leqslant \sum_{k=1}^{l}\left(\xi_{1}(k) \prod_{j=1}^{k-1} \xi_{3}(j)\right)+2 \Delta m, \tag{C.21}
\end{equation*}
$$

where the inequality follows since $\xi_{3}(k)$ is a convex combination of $\mathbb{P}\left(E_{4}\right)$ and $\mathbb{P}\left(E_{5}\right)$ (see (C.19)); hence $\xi_{3}(k) \leqslant 1 \forall k \in \mathbb{N}$. Now using (C.9),(C.10),(C.11),(C.12),(C.13) and Hoeffding's inequality [69] together with the fact that the rewards are bounded in $[0,1]$, we conclude

$$
\begin{array}{r}
\left\{\mathbb{P}\left(E_{1}\right), \mathbb{P}\left(E_{2}\right)\right\} \leqslant \exp \left(-(\Delta+\delta)^{2} m / 2\right) \\
\left\{\mathbb{P}\left(E_{3}\right), \mathbb{P}\left(E_{4}^{c}\right)\right\} \leqslant 2 \exp \left(-\delta^{2} m / 2\right) \\
\mathbb{P}\left(E_{5}\right) \leqslant \exp \left(-(\Delta-\delta)^{2} m / 2\right) \tag{C.24}
\end{array}
$$

From (C.17), (C.22) and (C.23), it follows that

$$
\begin{equation*}
\xi_{1}(k) \leqslant 2 \Delta \exp \left(-\delta^{2} m / 2\right) n+2 \Delta m \leqslant 2 \Delta+2 \Delta m \tag{C.25}
\end{equation*}
$$

where the last inequality follows since $m=\left\lceil\left(2 / \delta^{2}\right) \log n\right\rceil,\lceil\cdot\rceil$ being the "ceiling" operator. Using (C.21) and (C.25), we now have

$$
\begin{equation*}
\mathbb{E} R_{n}^{\pi} \leqslant 2 \Delta\left[1+\sum_{k=1}^{l} \prod_{j=1}^{k-1} \xi_{3}(j)\right](m+1) . \tag{C.26}
\end{equation*}
$$

From (C.19), observe that

$$
\begin{equation*}
\xi_{3}(k) \leqslant 1-\left(\alpha_{k}+\tilde{\alpha}_{k}-2 \alpha_{k} \tilde{\alpha}_{k}\right) \mathbb{P}\left(E_{5}^{c}\right) \leqslant \exp \left[-\left(\alpha_{k}+\tilde{\alpha}_{k}-2 \alpha_{k} \tilde{\alpha}_{k}\right) \mathbb{P}\left(E_{5}^{c}\right)\right] \quad \forall k \in \mathbb{N} \tag{C.27}
\end{equation*}
$$

where the last inequality follows since $1-x \leqslant \exp (-x)$. From (C.26) and (C.27), we obtain

$$
\mathbb{E} R_{n}^{\pi} \leqslant 2 \Delta\left[1+\sum_{k=1}^{l} \exp \left(-\mathbb{P}\left(E_{5}^{c}\right) \sum_{j=1}^{k-1}\left(\alpha_{j}+\tilde{\alpha}_{j}-2 \alpha_{j} \tilde{\alpha}_{j}\right)\right)\right](m+1) .
$$

Recall from (C.24) that $\mathbb{P}\left(E_{5}^{c}\right) \geqslant 1-\exp \left(-(\Delta-\delta)^{2} m / 2\right)$. Since $m=\left\lceil\left(2 / \delta^{2}\right) \log n\right\rceil$, it follows that $\exp \left(-(\Delta-\delta)^{2} m / 2\right)<1 / 2$ for $n>2\left(\frac{\delta}{\Delta-\delta}\right)^{2}$. Therefore, for $n$ large enough, we have

$$
\begin{align*}
\mathbb{E} R_{n}^{\pi} & \leqslant 2 \Delta\left[1+\sum_{k=1}^{l} \exp \left(-\frac{1}{2} \sum_{j=1}^{k-1}\left(\alpha_{j}+\tilde{\alpha}_{j}-2 \alpha_{j} \tilde{\alpha}_{j}\right)\right)\right](m+1) \\
& \leqslant 2 \Delta\left[3+\sum_{k=3}^{l} \exp \left(-\frac{1}{2} \sum_{j=1}^{k-1}\left(\alpha_{j}+\tilde{\alpha}_{j}-2 \alpha_{j} \tilde{\alpha}_{j}\right)\right)\right](m+1) \\
& \leqslant 2 \Delta\left[3+\sum_{k=3}^{l} \exp \left(-\frac{1}{2} \sum_{j=2}^{k-1}\left(\alpha_{j}+\tilde{\alpha}_{j}-2 \alpha_{j} \tilde{\alpha}_{j}\right)\right)\right](m+1) . \tag{C.28}
\end{align*}
$$

This concludes the basic analysis of Algorithm 1. We will use these results in subsequent sub-sections to provide the proofs for specific functional forms of $\alpha(t)$.

## C.3.1 Proof of Theorem 15

Recall that $\alpha_{j}:=\alpha((2 j-2) m+1)$. Since $\alpha(t) \sim c t^{-\gamma}$, it follows that for $n$ large enough (equivalently, $m$ large enough since $m=\left\lceil\left(2 / \delta^{2}\right) \log n\right\rceil$ ), we have $\alpha_{j} \leqslant 1 / 2$ for all $j \geqslant 2$, which implies $\alpha_{j}+\tilde{\alpha}_{j}-2 \alpha_{j} \tilde{\alpha}_{j} \geqslant \alpha_{j}$ for all $j \geqslant 2$. Therefore, it follows from (C.28) that for $n$ large enough,

$$
\begin{equation*}
\mathbb{E} R_{n}^{\pi} \leqslant 2 \Delta\left[3+\sum_{k=3}^{l} \exp \left(-\frac{1}{2} \sum_{j=2}^{k-1} \alpha_{j}\right)\right](m+1) . \tag{C.29}
\end{equation*}
$$

Now, since $\alpha(t) \sim c t^{-\gamma}$, it follows that for $n$ large enough (equivalently, $m$ large enough),

$$
\begin{equation*}
\alpha_{j} \geqslant \frac{c}{(2 j m)^{\gamma}}, \tag{C.30}
\end{equation*}
$$

Combining (C.29) and (C.30), we get that for $n$ large enough,

$$
\begin{align*}
\mathbb{E} R_{n}^{\pi} & \leqslant 2 \Delta\left[3+\sum_{k=3}^{l} \exp \left(-\frac{c}{m^{\gamma} 2^{\gamma+1}} \sum_{j=2}^{k-1} j^{-\gamma}\right)\right](m+1)  \tag{C.31}\\
& \leqslant 2 \Delta\left[3+\sum_{k=3}^{l} \exp \left(-\frac{c}{m^{\gamma} 2^{\gamma+1}} \int_{2}^{k} x^{-\gamma} d x\right)\right](m+1) \\
& =2 \Delta\left[3+\sum_{k=3}^{l} \exp \left(-\frac{c\left(k^{1-\gamma}-2^{1-\gamma}\right)}{(1-\gamma) m^{\gamma} 2^{\gamma+1}}\right)\right](m+1) \\
& \leqslant 2 \Delta\left[3+\sum_{k=3}^{l} \exp \left(-\frac{c\left(k^{1-\gamma}-2^{1-\gamma}\right)}{4(1-\gamma) m^{\gamma}}\right)\right](m+1) \\
& \leqslant 6 \Delta\left[1+\sum_{k=3}^{l} \exp \left(-\frac{c k^{1-\gamma}}{4(1-\gamma) m^{\gamma}}\right)\right](m+1) \tag{C.32}
\end{align*}
$$

where the last inequality holds since $m=\left\lceil\left(2 / \delta^{2}\right) \log n\right\rceil$ and $n$ is large enough. Now observe that

$$
\begin{align*}
\mathbb{E} R_{n}^{\pi} & \leqslant 6 \Delta\left[1+\sum_{k=3}^{l} \exp \left(-\frac{c k^{1-\gamma}}{4(1-\gamma) m^{\gamma}}\right)\right](m+1) \\
& \leqslant 6 \Delta\left[1+\int_{2}^{l} \exp \left(-\frac{c x^{1-\gamma}}{4(1-\gamma) m^{\gamma}}\right) d x\right](m+1) \\
& \leqslant 6 \Delta\left[1+\int_{2}^{n /(2 m)} \exp \left(-\frac{c x^{1-\gamma}}{4(1-\gamma) m^{\gamma}}\right) d x\right](m+1) \tag{C.33}
\end{align*}
$$

where the last inequality follows since $l=\lfloor n /(2 m)\rfloor$. We now focus on solving the integral. Define

$$
\begin{aligned}
\mathcal{I} & :=\int_{2}^{n /(2 m)} \exp \left(-\frac{c x^{1-\gamma}}{4(1-\gamma) m^{\gamma}}\right) d x \\
& \leqslant \int_{0}^{\infty} \exp \left(-\frac{c x^{1-\gamma}}{4(1-\gamma) m^{\gamma}}\right) d x \\
& =(1-\gamma)^{\frac{\gamma}{1-\gamma}}\left(\frac{4 m^{\gamma}}{c}\right)^{\frac{1}{1-\gamma}} \int_{0}^{\infty} z^{\frac{\gamma}{1-\gamma}} \exp (-z) d z \\
& \leqslant\left(\frac{4}{c}\right)^{\frac{1}{1-\gamma}}\left(\int_{0}^{\infty} z^{\frac{\gamma}{1-\gamma}} \exp (-z) d z\right) m^{\frac{\gamma}{1-\gamma}} \\
& \left.\leqslant\left(\frac{4}{c}\right)^{\frac{1}{1-\gamma}}\left(\int_{0}^{\infty} z^{\frac{\gamma}{1-\gamma}}\right] \exp (-z) d z\right) m^{\frac{\gamma}{1-\gamma}}
\end{aligned}
$$

where $(\ddagger)$ follows after the variable substitution $z=\frac{c x^{1-\gamma}}{4(1-\gamma) m^{\gamma}}$. Now, the $\left\lceil\frac{\gamma}{1-\gamma}\right\rceil^{\text {th }}$ moment of a unit rate exponential random variable is given by the factorial of $\left\lceil\frac{\gamma}{1-\gamma}\right\rceil$, denoted by $\mathfrak{F}\left(\left\lceil\frac{\gamma}{1-\gamma}\right\rceil\right)$. Thus, we have

$$
\begin{equation*}
I \leqslant\left(\frac{4}{c}\right)^{\frac{1}{1-\gamma}} \mathfrak{F}\left(\left[\frac{\gamma}{1-\gamma}\right]\right) m^{\frac{\gamma}{1-\gamma}} \tag{C.34}
\end{equation*}
$$

Combining (C.33) and (C.34), we conclude that for large enough $n$,

$$
\mathbb{E} R_{n}^{\pi} \leqslant 24 \Delta\left(\frac{4}{c}\right)^{\frac{1}{1-\gamma}} \mathfrak{F}\left(\left[\frac{\gamma}{1-\gamma}\right\rceil\right) m^{\frac{1}{1-\gamma}}
$$

Finally, since $m=\left\lceil\left(2 / \delta^{2}\right) \log n\right\rceil$, the stated assertion follows, i.e.,

$$
\limsup _{n \rightarrow \infty} \frac{\mathbb{E} R_{n}^{\pi}}{(\log n)^{\frac{1}{1-\gamma}}} \leqslant 24 \Delta\left(\frac{8}{\delta^{2} c}\right)^{\frac{1}{1-\gamma}} \mathfrak{F}\left(\left[\frac{\gamma}{1-\gamma}\right]\right) .
$$

## C.3.2 Proof of Theorem 17

Again, we pick things up from (C.28). We know that

$$
\mathbb{E} R_{n}^{\pi} \leqslant 2 \Delta\left[3+\sum_{k=3}^{l} \exp \left(-\frac{1}{2} \sum_{j=2}^{k-1}\left(\alpha_{j}+\tilde{\alpha}_{j}-2 \alpha_{j} \tilde{\alpha}_{j}\right)\right)\right](m+1),
$$

where $m=\left\lceil\left(2 / \delta^{2}\right) \log n\right\rceil, \alpha_{j}=\alpha((2 j-2) m+1)$ and $\tilde{\alpha}_{j}=\alpha((2 j-2) m+2)$. Since $\alpha(t)=$ $g\left(\mathcal{J}_{t-1}\right)$, we have $\alpha_{j}=g(2(j-1))$ and $\tilde{\alpha}_{j}=g(2 j-1)$. Since $g(\cdot) \leqslant c \leqslant 1 / 2$, it follows that $\alpha_{j}+\tilde{\alpha}_{j}-2 \alpha_{j} \tilde{\alpha}_{j} \geqslant \alpha_{j}=g(2(j-1))$ for all $j \geqslant 1$. Therefore, one has

$$
\begin{align*}
\mathbb{E} R_{n}^{\pi} & \leqslant 2 \Delta\left[3+\sum_{k=3}^{l} \exp \left(-\frac{1}{2} \sum_{j=2}^{k-1} g(2(j-1))\right)\right](m+1) \\
& =2 \Delta\left[3+\sum_{k=3}^{l} \exp \left(-\frac{c}{2} \sum_{j=2}^{k-1}(2 j-1)^{-\gamma}\right)\right](m+1) \\
& \leqslant 2 \Delta\left[3+\sum_{k=3}^{l} \exp \left(-\frac{c}{2^{\gamma+1}} \sum_{j=2}^{k-1} j^{-\gamma}\right)\right](m+1) . \tag{C.35}
\end{align*}
$$

Drawing upon structural similarities between (C.35) and (C.31), one can proceed along an analogous sequence of steps to eventually conclude

$$
\limsup _{n \rightarrow \infty} \frac{\mathbb{E} R_{n}^{\pi}}{\log n} \leqslant\left(\frac{48 \Delta}{\delta^{2}}\right)\left(\frac{4}{c}\right)^{\frac{1}{1-\gamma}} \mathfrak{F}\left(\left\lceil\frac{\gamma}{1-\gamma}\right\rceil\right),
$$

where $\mathfrak{F}(\cdot)$ denotes the Factorial function.

## C. 4 Auxiliary results used in the analysis of Algorithm 5

Fact 4 (Lemma 2 of [21]) Consider a stochastic two-armed bandit with [0, 1]-valued rewards and gap $\Delta$. Let $\left(X_{i, j}\right)_{j \geqslant 1}$ be the reward sequence associated with arm $i$ and $N_{i}(n)$ its sample count until time $n$ under UCB1 [10]. Define $M_{n}:=\min \left(N_{1}(n), N_{2}(n)\right)$ and consider the following stopping time:

$$
T:=\inf \left\{n \geqslant 2:\left|\sum_{j=1}^{M_{n}}\left(X_{1, j}-X_{2, j}\right)\right|<4 \sqrt{M_{n} \log M_{n}}\right\} .
$$

Then, the following results hold:

1. If $\Delta>0$, then $\mathbb{P}(T=\infty) \geqslant \beta_{\Delta}>0$, where $\beta_{\Delta}$ is as defined in (4.2).
2. If $\Delta=0$, then $\mathbb{E} T \leqslant C_{0}<\infty$, where $C_{0}$ is some absolute constant.

## C. 5 Analysis of Algorithm 5

Algorithm 5 runs in epochs of stochastic durations that are determined online in an adaptive manner. Let the sequence of epoch durations be $\left(T_{k}: k=1,2, \ldots\right)$. Define the episode count $W_{n}:=\inf \left\{l \in \mathbb{N}: \sum_{k=1}^{l} T_{k} \geqslant n\right\}$. Let $I_{k}(i)$ denote the event that each of the two arms queried at the beginning of epoch $k$ has type $i(i$ is an element of $\mathcal{K}=\{1,2\})$. Similarly, let $D_{k}$ denote the event that the aforementioned arms have "distinct" types. Suppose that $S_{n}$ denotes the cumulative pseudo-regret of UCB1 [10] after $n$ plays in an independent instance of a two-armed bandit with gap $\Delta$. Then, the evolution of the cumulative pseudo-regret of the policy $\pi$ given by Algorithm 5
after any number $n$ of plays satisfies

$$
\begin{aligned}
\tilde{R}_{n}^{\pi} & =\sum_{k=1}^{W_{n}-1}\left[\mathbb{1}\left\{D_{k}\right\} S_{T_{k}}+\mathbb{1}\left\{I_{k}(2)\right\} \Delta T_{k}\right]+\mathbb{1}\left\{D_{W_{n}}\right\} S_{\left(n-\sum_{k=1}^{W_{n}-1} T_{k}\right)}+\mathbb{1}\left\{I_{W_{n}}(2)\right\} \Delta\left(n-\sum_{k=1}^{W_{n}-1} T_{k}\right) \\
& \leqslant \sum_{(\dagger)}^{W_{n}-1}\left[\mathbb{1}\left\{D_{k}\right\} S_{n}+\mathbb{1}\left\{I_{k}(2)\right\} \Delta T_{k}\right]+\mathbb{1}\left\{D_{W_{n}}\right\} S_{n}+\mathbb{1}\left\{I_{W_{n}}(2)\right\} \Delta T_{W_{n}} \\
& =\sum_{k=1}^{W_{n}}\left[\mathbb{1}\left\{D_{k}\right\} S_{n}+\mathbb{1}\left\{I_{k}(2)\right\} \Delta T_{k}\right] \\
& =\sum_{k=1}^{\infty} \mathbb{1}\left\{W_{n} \geqslant k\right\} \mathbb{1}\left\{D_{k}\right\} S_{n}+\sum_{k=1}^{\infty} \mathbb{1}\left\{W_{n} \geqslant k\right\} \mathbb{1}\left\{I_{k}(2)\right\} \Delta T_{k},
\end{aligned}
$$

where $(\dagger)$ follows since the pseudo-regret of UCB1 is weakly increasing in the horizon, and $n-$ $\sum_{k=1}^{W_{n}-1} T_{k} \leqslant T_{W_{n}}$. Taking expectations and invoking Tonelli's theorem to interchange expectation and infinite-sum (since all summands are non-negative), we obtain

$$
\begin{equation*}
\mathbb{E} R_{n}^{\pi}=\mathbb{E} \tilde{R}_{n}^{\pi} \leqslant \mathbb{E} S_{n} \sum_{k=1}^{\infty} \mathbb{P}\left(D_{k}, W_{n} \geqslant k\right)+\Delta \sum_{k=1}^{\infty} \mathbb{P}\left(I_{k}(2), W_{n} \geqslant k\right) \mathbb{E}\left[T_{k} \mid I_{k}(2), W_{n} \geqslant k\right] . \tag{C.36}
\end{equation*}
$$

(Note that $\mathbb{E}\left[S_{n} \mid W_{n} \geqslant k, D_{k}\right]=\mathbb{E} S_{n}$ since $S_{n}$, by definition, is independent of $W_{n}$ and $D_{k}$.)

Lemma 9 (Bounded epochs for identical-typed arms) The following holds for any $k, n \in \mathbb{N}$ and arm-type $i \in\{1,2\}$ :

$$
\mathbb{E}\left[T_{k} \mid I_{k}(i), W_{n} \geqslant k\right]=\mathbb{E}\left[T_{k} \mid I_{k}(i)\right] \leqslant C_{0}<\infty,
$$

where $C_{0}$ is as given in Fact 4.

$$
\begin{aligned}
& \mathbb{E}\left[T_{k} \mid I_{k}(i)\right] \\
= & \mathbb{E}\left[T_{k} \mid I_{k}(i), W_{n} \geqslant k\right] \mathbb{P}\left(W_{n} \geqslant k \mid I_{k}(i)\right)+\mathbb{E}\left[T_{k} \mid I_{k}(i), W_{n}<k\right] \mathbb{P}\left(W_{n}<k \mid I_{k}(i)\right) \\
= & \mathbb{E}\left[T_{k} \mid I_{k}(i), W_{n} \geqslant k\right] \mathbb{P}\left(W_{n} \geqslant k \mid I_{k}(i)\right)+\mathbb{E}\left[T_{k} \mid I_{k}(i)\right] \mathbb{P}\left(W_{n}<k \mid I_{k}(i)\right),
\end{aligned}
$$

where $(\dagger)$ follows since $T_{k}$ is independent of $W_{n}$, given $I_{k}(i)$ and $k>W_{n}$. Thus,

$$
\mathbb{E}\left[T_{k} \mid I_{k}(i), W_{n} \geqslant k\right]=\mathbb{E}\left[T_{k} \mid I_{k}(i)\right] \leqslant C_{0},
$$

with the last inequality following from Fact 4.2.
Now coming back to the analysis of Algorithm 5, observe that from (C.36) and Lemma 9, one has

$$
\begin{equation*}
\mathbb{E} R_{n}^{\pi} \leqslant \mathbb{E} S_{n} \sum_{k=1}^{\infty} \mathbb{P}\left(D_{k}, W_{n} \geqslant k\right)+C_{0} \Delta \sum_{k=1}^{\infty} \mathbb{P}\left(I_{k}(2), W_{n} \geqslant k\right) \tag{C.37}
\end{equation*}
$$

Note that

$$
\begin{align*}
\mathbb{P}\left(W_{n} \geqslant k\right) & =\mathbb{P}\left(\sum_{m=1}^{k-1} T_{m}<n\right) \\
& =\mathbb{P}\left(\bigcap_{j=1}^{k-1}\left\{\sum_{m=1}^{j} T_{m}<n\right\}\right) \\
& =\mathbb{P}\left(T_{1}<n\right) \prod_{j=2}^{k-1} \mathbb{P}\left(\sum_{m=1}^{j} T_{m}<n \mid T_{1}<n, \ldots, \sum_{m=1}^{j-1} T_{m}<n\right) \\
& =\mathbb{P}\left(T_{1}<n\right) \prod_{j=2}^{k-1} \mathbb{P}\left(\sum_{m=1}^{j} T_{m}<n \mid \sum_{m=1}^{j-1} T_{m}<n\right) \\
& \leqslant \mathbb{P}\left(T_{1}<\infty\right) \prod_{j=2}^{k-1} \mathbb{P}\left(T_{j}<\infty \mid \sum_{m=1}^{j-1} T_{m}<n\right) . \tag{C.38}
\end{align*}
$$

This concludes the basic analysis of Algorithm 5. We will use these results in subsequent
sub-sections to prove guarantees under specific functional forms of $\alpha(t)$.

## C.5.1 Proof of Theorem 16

In this setting, $(\alpha(t): t=1,2, \ldots)$ is a non-increasing process. Observe that conditional on the event $E_{l}:=\left\{\sum_{m=1}^{j-1} T_{m}=l\right\}$, where $l<n$ is arbitrary, the probability that $T_{j}<\infty$ (for $j \geqslant 2$ ), satisfies

$$
\begin{aligned}
\mathbb{P}\left(T_{j}<\infty \mid E_{l}\right) & =\mathbb{P}\left(T_{j}<\infty \mid E_{l}, I_{j}(1) \cup I_{j}(2)\right) \mathbb{P}\left(I_{j}(1) \cup I_{j}(2) \mid E_{l}\right)+\mathbb{P}\left(T_{j}<\infty \mid E_{l}, D_{j}\right) \mathbb{P}\left(D_{j} \mid E_{l}\right) \\
& \underset{(\dagger)}{\leq} \mathbb{P}\left(I_{j}(1) \cup I_{j}(2) \mid E_{l}\right)+\mathbb{P}\left(T_{j}<\infty \mid D_{j}\right) \mathbb{P}\left(D_{j} \mid E_{l}\right) \\
& \underset{(\underset{+}{*})}{\mathbb{P}}\left(I_{j}(1) \cup I_{j}(2) \mid E_{l}\right)+\left(1-\beta_{\Delta}\right) \mathbb{P}\left(D_{j} \mid E_{l}\right) \\
& =1-\mathbb{P}\left(D_{j} \mid E_{l}\right) \beta_{\Delta}, \\
& =1-[\alpha(l+1)(1-\alpha(l+2))+\alpha(l+2)(1-\alpha(l+1))] \beta_{\Delta} \\
& \leqslant 1-\alpha(l+1) \beta_{\Delta} \\
& \underset{(*)}{\leqslant} 1-\alpha(n) \beta_{\Delta},
\end{aligned}
$$

where $(\dagger)$ follows because $T_{j}$ is independent of $E_{l}$, given $D_{j}$, and $(\ddagger)$ follows using Fact 4.1. Next, $(\star)$ follows since $\alpha(l+1) \leqslant 1 / 2$ by assumption and finally, $(*)$ since $l+1 \leqslant n$ and $\alpha(\cdot)$ is nonincreasing. Notice that although $(*)$ holds for $j \geqslant 2$, the same upper bound of $1-\alpha(n) \beta_{\Delta}$ holds trivially also for $\mathbb{P}\left(T_{1}<\infty\right)$ (proof is almost identical to that for $j \geqslant 2$ except that the probabilities are unconditional). Using (*) and said observation in (C.38), one concludes that

$$
\begin{equation*}
\mathbb{P}\left(W_{n} \geqslant k\right) \leqslant\left(1-\alpha(n) \beta_{\Delta}\right)^{k-1} . \tag{C.39}
\end{equation*}
$$

Now, we know from (C.37) that

$$
\mathbb{E} R_{n}^{\pi} \leqslant \mathbb{E} S_{n} \sum_{k=1}^{\infty} \mathbb{P}\left(W_{n} \geqslant k\right)+C_{0} \Delta \sum_{k=1}^{\infty} \mathbb{P}\left(W_{n} \geqslant k\right) \leqslant \frac{\mathbb{E} S_{n}+C_{0} \Delta}{\alpha(n) \beta_{\Delta}},
$$

where the last step follows using (C.39). Finally, using $\mathbb{E} S_{n} \leqslant(8 / \Delta) \log n+\left(1+\pi^{2} / 3\right) \Delta[10]$ and taking appropriate limits, the stated assertion follows.

## C.5.2 Proof of Theorem 18

In this setting, $\alpha(t)=g\left(\mathcal{J}_{t-1}\right)$, where $\mathcal{J}_{t-1}$ is the number of reservoir queries until time $t-$ 1 (inclusive) and $g(\cdot)$ is non-increasing. Since the dependence on $t$ is only through $\mathcal{J}_{t-1}$, joint probabilities in (C.37) decouple into products and we have the following lemma:

Lemma 10 (Product-form probabilities) Consider the following cases:

1. For any $k \in \mathbb{N}$ and any arm-type $i \in\{1,2\}$, the events $D_{k}, I_{k}(i)$ are independent of the "starting time" of epoch $k$.
2. For any $k \in \mathbb{N}$ and any arm-type $i \in\{1,2\}$, the events $D_{k}, I_{k}(i)$ depend on the "starting time" of epoch $k$ only through $k$.

In either case, one has for any $k, n \in \mathbb{N}$ and any arm-type $i \in\{1,2\}$ that

$$
\begin{aligned}
& \mathbb{P}\left(D_{k}, W_{n} \geqslant k\right)=\mathbb{P}\left(D_{k}\right) \mathbb{P}\left(W_{n} \geqslant k\right), \\
& \mathbb{P}\left(I_{k}(i), W_{n} \geqslant k\right)=\mathbb{P}\left(I_{k}(i)\right) \mathbb{P}\left(W_{n} \geqslant k\right) .
\end{aligned}
$$

Proof of Lemma 10. We know that

$$
\begin{aligned}
\mathbb{P}\left(D_{k}, W_{n}<k\right) & =\mathbb{P}\left(W_{n}<k\right) \mathbb{P}\left(D_{k} \mid k>W_{n}\right) \\
& =\mathbb{P}\left(W_{n}<k\right) \mathbb{P}\left(D_{k}\right) \\
& =\mathbb{P}\left(D_{k}\right)-\mathbb{P}\left(D_{k}\right) \mathbb{P}\left(W_{n} \geqslant k\right) \\
\Longrightarrow \mathbb{P}\left(D_{k}, W_{n} \geqslant k\right) & =\mathbb{P}\left(D_{k}\right) \mathbb{P}\left(W_{n} \geqslant k\right)
\end{aligned}
$$

The result for $I_{k}(i)$ can also be shown similarly.

Now coming back to the proof of Theorem 18, we will pick things up from (C.37). Using Lemma 10 with (C.37), we obtain

$$
\begin{align*}
\mathbb{E} R_{n}^{\pi} & \leqslant \mathbb{E} S_{n} \sum_{k=1}^{\infty} \mathbb{P}\left(D_{k}\right) \mathbb{P}\left(W_{n} \geqslant k\right)+C_{0} \Delta \sum_{k=1}^{\infty} \mathbb{P}\left(I_{k}(2)\right) \mathbb{P}\left(W_{n} \geqslant k\right) \\
& \leqslant \mathbb{E} S_{n} \sum_{k=1}^{\infty}[g(2 k-2)+g(2 k-1)-2 g(2 k-2) g(2 k-1)] \mathbb{P}\left(W_{n} \geqslant k\right)+C_{0} \Delta \sum_{k=1}^{\infty} \mathbb{P}\left(W_{n} \geqslant k\right) \\
& \leqslant 2 \mathbb{E} S_{n} \sum_{k=1}^{\infty} g(2 k-2) \mathbb{P}\left(W_{n} \geqslant k\right)+C_{0} \Delta \sum_{k=1}^{\infty} \mathbb{P}\left(W_{n} \geqslant k\right) \\
& =2 \mathbb{E} S_{n} \sum_{k=0}^{\infty} g(2 k) \mathbb{P}\left(W_{n} \geqslant k+1\right)+C_{0} \Delta \sum_{k=1}^{\infty} \mathbb{P}\left(W_{n} \geqslant k\right) \tag{C.40}
\end{align*}
$$

where (\$) follows since $g(\cdot)$ is non-increasing. Next, we upper bound $\mathbb{P}\left(W_{n} \geqslant k\right)$. Observe that conditional on the event $E_{l}:=\left\{\sum_{m=1}^{j-1} T_{m}=l\right\}$, where $l<n$ is arbitrary, the probability that $T_{j}<\infty$ (for $j \geqslant 2$ ), satisfies

$$
\begin{aligned}
\mathbb{P}\left(T_{j}<\infty \mid E_{l}\right) & =\mathbb{P}\left(T_{j}<\infty \mid E_{l}, I_{j}(1) \cup I_{j}(2)\right) \mathbb{P}\left(I_{j}(1) \cup I_{j}(2) \mid E_{l}\right)+\mathbb{P}\left(T_{j}<\infty \mid E_{l}, D_{j}\right) \mathbb{P}\left(D_{j} \mid E_{l}\right) \\
& \leqslant \mathbb{P}\left(I_{j}(1) \cup I_{j}(2) \mid E_{l}\right)+\mathbb{P}\left(T_{j}<\infty \mid D_{j}\right) \mathbb{P}\left(D_{j} \mid E_{l}\right) \\
& \underset{(\ddagger)}{\leq} \mathbb{P}\left(I_{j}(1) \cup I_{j}(2) \mid E_{l}\right)+\left(1-\beta_{\Delta}\right) \mathbb{P}\left(D_{j} \mid E_{l}\right) \\
& =1-\mathbb{P}\left(D_{j} \mid E_{l}\right) \beta_{\Delta}, \\
& =1-[\alpha(l+1)(1-\alpha(l+2))+\alpha(l+2)(1-\alpha(l+1))] \beta_{\Delta} \\
& \leq 1-\alpha(l+1) \beta_{\Delta} \\
& =(\star) \\
& =1-g(2 j-2) \beta_{\Delta},
\end{aligned}
$$

where $(\dagger)$ follows because $T_{j}$ is independent of $E_{l}$, given $D_{j}$, and ( $\ddagger$ ) follows using Fact 4.1. Next, $(\star)$ follows since $\alpha(l+1) \leqslant 1 / 2$ by assumption and finally, $(*)$ holds since $\alpha(l+1)=$ $g\left(\mathcal{J}_{l}\right)=g(2 j-2)$ on $E_{l}$. Notice that although $(*)$ holds for $j \geqslant 2$, the same upper bound of $1-g(2 j-2) \beta_{\Delta}$ holds trivially also for $\mathbb{P}\left(T_{1}<\infty\right)$ (proof is almost identical to that for $j \geqslant 2$
except that the probabilities are unconditional). Using (*) and said observation in (C.38), one concludes that

$$
\begin{equation*}
\mathbb{P}\left(W_{n} \geqslant k\right) \leqslant \prod_{j=1}^{k-1}\left(1-g(2 j-2) \beta_{\Delta}\right)=\prod_{j=0}^{k-2}\left(1-\beta_{\Delta} g(2 j)\right) . \tag{C.41}
\end{equation*}
$$

Combining (C.40) and (C.41), one obtains

$$
\begin{aligned}
\mathbb{E} R_{n}^{\pi} & \leqslant 2 \mathbb{E} S_{n} \sum_{k=0}^{\infty} g(2 k) \prod_{j=0}^{k-1}\left(1-\beta_{\Delta} g(2 j)\right)+C_{0} \Delta \sum_{k=0}^{\infty} \prod_{j=0}^{k-1}\left(1-\beta_{\Delta} g(2 j)\right) \\
& \leqslant\left(2 c \mathbb{E} S_{n}+C_{0} \Delta\right) \sum_{k=0}^{\infty} \prod_{j=0}^{k-1}\left(1-\beta_{\Delta} g(2 j)\right) \\
& \leqslant\left(2 c \mathbb{E} S_{n}+C_{0} \Delta\right) \sum_{k=0}^{\infty} \exp \left(-\beta_{\Delta} \sum_{j=0}^{k-1} g(2 j)\right)
\end{aligned}
$$

where $\left(\#_{1}\right)$ follows since $g(\cdot)$ is non-increasing with $g(0)=c$, and $\left(\#_{2}\right)$ follows using the identity $\log (1+x) \leqslant x \forall x>-1$. Finally, using $\mathbb{E} S_{n} \leqslant(8 / \Delta) \log n+\left(1+\pi^{2} / 3\right) \Delta$ [10] and taking appropriate limits, the stated assertion follows.


[^0]:    ${ }^{1}$ This is a standard technique for performance evaluation of stochastic systems commonly used in the operations research and mathematics literature, see, e.g., [22].
    ${ }^{2}$ Assumed to be vanishing in $n$; standard versions of the algorithm involve fixed prior variances.

[^1]:    ${ }^{3}$ These assumptions can be relaxed in the spirit of [10]; our results also extend to sub-Gaussian rewards.

[^2]:    ${ }^{4}$ This is the version of Thompson Sampling that is based on Gaussian priors and Gaussian likelihoods, not the classical Beta-Bernoulli version that has a minimax regret of $O(\sqrt{n \log n})$ [7].

[^3]:    ${ }^{5}$ A first principles-based approach that applies directly (without triangulation limits) to the case of non-vanishing prior variances is provided in Theorem 6.2 of [29].

[^4]:    ${ }^{6}$ Previous best result for UCB was an upper bound of $O(\sqrt{n \log n})$ [2]. It remained an open problem whether this bound was, in fact, achieved in the worst-case setting; Theorem 4 answers this affirmatively.

