A qualitative and quantitative analysis of the bias caused by adaptivity in multi-armed bandits

Jaehyeok Shin

May 2020

Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Aaditya Ramdas (Co-Chair)
Alessandro Rinaldo (Co-Chair)
Larry Wasserman
Alexander (Sasha) Rakhlin (MIT, Boston)
Gábor Lugosi (UPF, Barcelona)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Keywords: Bias, Multi-armed bandits, Mean estimation, Adaptive data analysis.

To my family

Abstract

In classical and non-adaptive data analysis, the target of interest is typically fixed in advance, and a fixed number of samples are collected in an i.i.d. manner to conduct statistical inference on the target. In many cases, however, data are collected and analyzed adaptively. For example, in the multi-armed bandit setting, data are collected sequentially and adaptively such that at every round, a sample is drawn from an arm which is selected based on the sampling history so far. The sampling procedure can be stopped based on a data-driven stopping rule. Furthermore, the adaptively collected data are often used to identify an interesting target and the same data are used to conduct statistical inference on this target. Even though this kind of adaptive scheme is prevalent in data analysis, theoretical justifications for commonly used inference procedures are not yet sufficiently developed. This disparity challenges the validity of decisions we make based on adaptive data analyses. In order to close the gap, the thesis first focuses on the mean estimation problem for multi-armed bandits. We derive a qualitative characterization of the adaptive mean estimation procedure which determines the sign of bias of the sample mean for each arm. We provide sharp bounds on both the bias and the risk which show that even though the sample mean is biased under adaptive schemes, the size of the risk is as small as one can achieve under the non-adaptive i.i.d. setting.

Acknowledgments

First, I wish to express my sincere appreciation to my advisors, Aaditya Ramdas and Alessandro Rinaldo. None of my works have been accomplished without their endless supports and insightful advice. During my PhD program, they have been my best role models for being both professional researcher and delightful collaborator which makes our every meeting full of joy.

I also would like to express my special thanks to Larry Wasserman who I most look up to as a great scholar and teacher. If someone asks me who I want to become like, I will not hesitate to say I hope to be like him in the future. His office has always been opened to me, and every time I visited his office, I was able to have a lesson from him.

I wish to show my gratitude to my other committee members, Sasha Rakhlin and Gábor Lugosi for giving me valuable feedback and comments on my thesis. Their insights were crucial to set the direction of this work properly.

The CMU Statistics & Data Science Department has a wonderful team of faculties who have been my teachers, senior researchers, and friends. Being a student of them is one of the most grateful experiences in my life. I especially thank to Valerie Ventura and Ann Lee who taught me how to convert my statistical knowledge into real data analyses.

I cannot thank our department staff enough. Their professional supports have been an essential part of my student life. They always make me feel like I have a strong team of supporters, which makes able me concentrate on my research without any concerns.

Thanks to my awesome colleagues and friends in our department. Most of my memorable memories in CMU consist of delightful conversations with them in FMS. I especially thank our cohort members, Ilmun Kim, Kayla Frisoli, Neil Spencer, Alden Green, Xiao Hui Tai, Daren Wang, Zongge Liu, and Octavio Mesner for being always with me together. I also would like to thank our SAC members who have always been sincere advocators of me and other students.

I would like to pay my special regards to the members of my research groups. Especially, I thank Jisu Kim for having been my best collaborator and friend. I also wish to thank Boyan Duan, Jinjin Tian, Tudor Manole, Pratik Patil, Ian Waudby-Smith, Robin Dunn, Ojash Neopane, Chirag Gupta and Eugene Katsevich for their valuable feedback on my works.

Many Korean communities in Pittsburgh make my life in Pittsburgh as in a second hometown. I especially thank my Korean gang in our department, Jisu Kim, Sangwon Hyun, Ilmun Kim, YJ Choe, Kwangho Kim, Heejong Bong, and Beomjo Park.

The department of Statistics in SNU is my intellectual hometown in which I started to dream of being a statistician. I would like to show my special thanks to Prof. Byeong U. Park, Myunghee Cho Paik, Yongdai Kim, Woncheol Jang, Taesung Park, and Hee-Seok Oh for being my great teachers and supporters.

Without endless supports and prayers from my family, I would not have even started to think about becoming a researcher. My parents have been shows theirs endless love and forgiving to their son which always gives me the courage to start all over again whenever I failed.

Finally, to my wife, Jiyong, thanks for having been with me. Nothing would have done without your great love and dedication. You are my true hero.

Contents

Chapter 1

Introduction

In many real-world settings, data are collected in an adaptive manner from several distributions (arms), as captured by the classic stochastic multi-armed bandits (MAB) framework [Robbins, 1952]. The data collection procedure (henceforth, algorithm) may have been primarily designed for purposes such as testing a hypothesis, minimizing regret or identifying the best arm. In each round, the algorithm draws a sample from one of the arms based on the previously observed data (*adaptive sampling*). The algorithm may also be terminated based on a data-driven stopping rule rather than at a fixed time (*adaptive stopping*).

Even though mean estimation may not have been the primary objective, the sample means of arms might nevertheless be calculated later on. For example, after identifying the best arm, it is natural to want an estimate of its mean reward. In "off-policy evaluation" [Li et al., 2015], mean reward estimates from a current policy are used to evaluate the performance of a different policy before actually implementing the latter. An analyst can choose a specific target arm based on the collected data (*adaptive choosing*), for example focusing on certain "promising" arms. Furthermore, the analyst may wish to analyze the data at some past times, as if the experiment had stopped earlier (*adaptive rewinding*).

Among several possible mean estimators, we focus on the sample mean which is arguably the simplest and most commonly used in practice. In the classical nonadaptive setting, the sample mean has favorable properties. In particular, it is unbiased, consistent, and converges almost surely to the true mean, μ. Additionally, under tail assumptions such as sub-Gaussian or sub-exponential conditions, the sample mean is tightly concentrated around μ. Lastly, the sample mean has minimax optimal risk with respect to suitable loss functions such as the ℓ_2 loss for distributions with a finite variance and the Kullback-Leibler (KL) loss for distributions in a natural exponential family.

In this thesis, we study the bias, risk and consistency of sample means under all four aforementioned notions of adaptivity (henceforth called the "fully adaptive setting"). As a qualitative analysis, we first derive a characterization of the adaptive mean estimation procedure which determines the sign of the unconditional bias of the sample mean, i.e., the bias obtained by accounting for all possible outcomes of the MAB experiment (Chapter 2). Then, we extend it to the conditional bias case in which the sample means of the arms are counted only when certain outcomes have occurred (Chapter 3). As a quantitative analysis, we provide sharp bounds on both the bias and the risk which show that even though the sample mean is biased under adaptive schemes, the size of the risk is as small as one can achieve under the non-adaptive i.i.d. setting which also demonstrates that the size of bias is small(Chapter 4). These results are based on a recent publication and preprints [Shin et al., 2019a,b, 2020].

Before presenting accomplished results in this thesis , we briefly introduce the stochastic multi-armed model and formalize the four notions of adaptivity.

1.1 The stochastic multi-armed bandit model

Let $P_1, \ldots, P_K$ be K distributions of K arms with finite means $\mu_k = \int x \mathrm{d}P_k(x)$. If it is clear from context, every inequality and equality between two random variables is understood in the almost sure sense.

1.1.1 Four sources of adaptivity in MABs

Data collection and inference procedures based on stochastic MAB algorithms consist of the following sampling, stopping, choosing and rewinding rules:

- **Sampling:** For each time $t \geq 1$, choose a possibly randomized index of arm A_t from a multinomial distribution $\mathrm{Multi}(1, \nu_t) \in \{0, 1, \ldots, K\}$. Then, we draw a sample (also called a reward) Y_t from the chosen arm P_{A_t}. Here, $\nu_t \in [0,1]^K$ refers to conditional probabilities of sampling each arm defined by

$$\nu_t(k) := \nu_t(k \mid \mathcal{D}_{t-1}) \in [0,1] \ \ \forall k \in [K],$$

  with the constraint $\sum_{k=1}^K \nu_t(k) = 1$, and $\mathcal{D}_{t-1}$ refers to observed data up to time $t-1$ including all possible external random sources. If each ν_t does not depend on the previous data $\mathcal{D}_{t-1}$, we call it a nonadaptive sampling rule.

  With a proper choice of ν_t, this sampling rule captures a broad class of existing methods including fixed design, random allocation, ϵ-greedy [Sutton and Barto, 1998], upper confidence bound (UCB) algorithms [Auer et al., 2002, Audibert and Bubeck, 2009, Garivier and Cappé, 2011, Kalyanakrishnan et al., 2012, Jamieson et al., 2014] and Thompson sampling [Thompson, 1933, Agrawal and Goyal, 2012, Kaufmann et al., 2012]. To be specific, ϵ-greedy, UCB, and Thomson sampling have the following sampling rules.

  - ϵ-greedy algorithm : For any $k \in [K]$ and $t \in [\mathcal{T}]$,

$$\nu_t(k) = \begin{cases} 1-\epsilon & \text{if } k = \arg\max_{j \in [K]} \widehat{\mu}_j(t-1), \\ \frac{\epsilon}{K-1} & \text{otherwise.} \end{cases}$$

  - UCB : For any $k \in [K]$ and $t \in [\mathcal{T}]$,

$$\nu_t(k) = \begin{cases} 1 & \text{if } k = \arg\max_{j \in [K]} \widehat{\mu}_j(t-1) + u_{t-1}(S_j(t-1), N_j(t-1)), \\ 0 & \text{otherwise,} \end{cases}$$

    where $(s,n) \mapsto u_{t-1}(s,n)$ is a non-negative function which is increasing and decreasing with respect to the first and second inputs respectively for each t. For example, a simple version of UCB uses $u_{t-1}(s,n) = \sqrt{\frac{2\log(1/\delta)}{n}}$ for a properly chosen constant $\delta \in (0,1)$.

  - Thompson sampling : For any $k \in [K]$ and $t \in [\mathcal{T}]$,

$$\nu_t(k) \propto \pi(k = \arg\max_j \mu_j \mid A_1, Y_1, \ldots, A_{t-1}, Y_{t-1}).$$

    where π is a prior on $(\mu_1, \ldots, \mu_K)$ or, more generally, on parameters of arms $(\theta_1, \ldots, \theta_K)$. In particular, if underlying arms are Gaussian with common variance σ^2 and if we impose independent Gaussian prior $N(\mu_{k,0}, \sigma_0^2)$ on each arm k, the corresponding Thompson sampling is statistically equivalent to the following rule.

$$\nu_t(k) = \begin{cases} 1 & \text{if } k = \arg\max_{j \in [K]} \widetilde{\mu}_j(t-1) + \sigma_j(t-1) Z_{j,t-1} \\ 0 & \text{otherwise,} \end{cases}$$

where each $Z_{j,t-1}$ is an independent draw from $N(0,1)$ and $\widetilde{\mu}_j(t-1), \sigma_k(t-1)$ are the posterior mean and standard deviation of arm j, given as

$$\widetilde{\mu}_j(t-1) = \frac{\mu_{j,0}/\sigma_0^2 + N_j(t-1)\widehat{\mu}_j(t-1)/\sigma^2}{1/\sigma_0^2 + N_j(t-1)/\sigma^2}, \;\; \sigma_j(t-1) = \left(1/\sigma_0^2 + N_j(t-1)/\sigma^2\right)^{-1/2}.$$

With a slight abuse of notation, we denote with $\nu = \{\nu_t\}_{t\geq 1}$ the sampling rule. If, for each t, the sampling rule ν_t is independent of the arm realization observed so far, namely $Y_1, \ldots, Y_{t-1}$ (but not necessarily of the sampling history $A_1, \ldots, A_{t-1}$), we call it a *nonadaptive* sampling rule.

- **Stopping:** Given a sampling rule, let $\{\mathcal{F}_t\}$ be a filtration such that each A_t is $\mathcal{F}_{t-1}$-measurable and each Y_t and $\mathcal{D}_t$ are $\mathcal{F}_t$-measurable. Let $\mathcal{T}$ be a stopping time with respect to the filtration $\{\mathcal{F}_t\}$. Then, for each time t, the stopping event $\{\mathcal{T} = t\}$ is $\mathcal{F}_t$-measurable which can be used to characterize a stopping rule of a MAB algorithm.

  For example, the following stopping time corresponds to the stopping rule in which we stop whenever the absolute difference between sample means for arm 1 and arm 2 exceeding a fixed threshold $\delta > 0$:

$$\mathcal{T} := \inf\left\{t \geq 1 : |\widehat{\mu}_1(t) - \widehat{\mu}_2(t)| > \delta\right\}.$$

  Here, we refer $\widehat{\mu}_k(t)$ to the sample mean for arm k defined as

$$\widehat{\mu}_k(t) := \frac{S_k(t)}{N_k(t)}, \tag{1.1}$$

  where $S_k(t)$ and $N_k(t)$ are the running sum and number of draws for arm k defined respectively $S_k(t) := \sum_{s=1}^t \mathbb{1}(A_s = k)Y_s$ and $N_k(t) := \sum_{s=1}^t \mathbb{1}(A_s = k)$ for each $k \in [K]$ and $t \geq 1$.[1] Note that, by the construction, $S_k(t)$, $\widehat{\mu}_k(t)$ are $\mathcal{F}_t$-measurable and $N_k(t)$ is $\mathcal{F}_{t-1}$-measurable for each $k \in [K]$ and $t \geq 1$.

  If a stopping rule is nonadaptive, that is, if the corresponding stopping time is $\mathcal{F}_0$-measurable, we use T instead of $\mathcal{T}$ to refer the stopping time which includes all fixed time and, more generally, all data-independent stopping rules.

- **Choosing:** After stopping, let $\mathcal{D}_\mathcal{T}$ denote collected data up to the stopping time $\mathcal{T}$. Based on $\mathcal{D}_\mathcal{T}$ with possible randomization, we choose a data-dependent arm based on a choosing rule $\kappa : \mathcal{D}_\mathcal{T} \mapsto [K]$. In this thesis, we denote the index of chosen arm $\kappa(\mathcal{D}_\mathcal{T})$ by κ for simplicity.

  For example, in the best arm identification, we often choose the best arm as the arm with the largest sample mean at the stopping time $\mathcal{T}$. In this case, the choosing rule is given as $\kappa = \arg\max_{k\in[K]} \widehat{\mu}_k(\mathcal{T})$. If κ does not depend on the observed data, we call it a nonadaptive choosing rule.

- **Rewinding:** Optionally, we may adaptively rewind the clock to focus on a previous random time $\tau \leq \mathcal{T}$ to characterize the past behavior of a chosen sample mean $\widehat{\mu}_\kappa(\tau)$. The rewound time τ is assumed to be measurable with respect to $\mathcal{F}_\mathcal{T}$; in particular, it is *not* a stopping time. For instance, $\tau = \arg\max_{t\leq\mathcal{T}} \widehat{\mu}_\kappa(t)$ is a rewound time. We may care about the bias of the sample mean at this "extreme" time τ. If we do not rewind, then $\tau = \mathcal{T}$.

The phrase "fully adaptive setting" refers to the scenario of running an adaptive sampling algorithm until an adaptive stopping time $\mathcal{T}$, and asking about the sample mean of an adaptively chosen arm κ at an adaptively rewound time τ. When we are not in the fully adaptive setting, we explicitly mention what aspects are adaptive.

[1] Throughout this thesis, we assume $N_k(t) \geq 1$ whenever we make a statement about the sample mean $\widehat{\mu}_k(t)$.

1.1.2 The tabular model of stochastic MABs

To analyze the theoretical properties of a MAB experiment, it is convenient to express it using an infinite tabular representation of the arm rewards. In the tabular model, we assume that each observation Y_t comes from an imaginary $\mathbb{N} \times K$ table or stacks of samples $\{X^*_{i,k}\}_{i\in\mathbb{N},k\in[K]} := X^*_\infty$ where each $X^*_{i,k}$ is an independent draw from P_k, and the observation Y_t is equal to $X^*_{i,k}$ if $N_k(t) = i$ and $A_t = k$. Here, we put an asterisk to clarify that it is counterfactual and not necessarily observable.

Now, given a MAB algorithm, let W_t be a random variable which accounts for all possible external randomness of the algorithm for each time t. Next, we define $\mathcal{D}^*_\infty = X^*_\infty \cup \{W_t\}_{t=0}^\infty$ as the collection of all possible arm rewards and external randomness. Then, given $\mathcal{D}^*_\infty$, every property and outcome of a MAB experiment becomes a deterministic function of $\mathcal{D}^*_\infty$. In particular $\mathcal{T}$, κ and $N_k(t)$, for each t and k, can be expressed as some functions of $\mathcal{D}^*_\infty$. For instance, given the above tabular MAB setup (which is statistically indistinguishable from the setup described in the previous subsection), one may then find deterministic functions $f_{t,k}$ and f^*_k such that

$$N_k(\mathcal{T}) = \sum_{t\geq 1} \underbrace{\mathbb{1}(A_t = k)\,\mathbb{1}(\mathcal{T} \geq t)}_{\mathcal{F}_{t-1}\text{-measurable}} = \sum_{t\geq 1} f_{t,k}(\mathcal{D}_{t-1}) \equiv f^*_k(\mathcal{D}^*_\infty). \tag{1.2}$$

Specifically, the function $f_{t,k}(\cdot)$ evaluates to one if and only if we do not stop at time $t-1$, and pull arm k at time t. Indeed, given $\mathcal{D}^*_\infty$, the stopping time $\mathcal{T}$ is deterministic and so is the number of times $N_k(\mathcal{T})$ that a fixed arm k is pulled, and this is what f^*_k captures. Along the same lines, the number of draws from a chosen arm κ at stopping time $\mathcal{T}$ can be written in terms of the tabular data as

$$N_\kappa(\mathcal{T}) \;=\; \sum_{k=1}^K \mathbb{1}(\kappa = k)\, N_k(\mathcal{T}) \;\equiv\; \sum_{k=1}^k g^*_k(\mathcal{D}^*_\infty) f^*_k(\mathcal{D}^*_\infty) \tag{1.3}$$

for some deterministic set of functions $\{g^*_k\}$. Indeed, g^*_k evaluates to one if after stopping, we choose arm k, which is a fully deterministic choice given $\mathcal{D}^*_\infty$.

1.2 What the bias means and why the sample mean is based in MABs

We finish the introduction by defining the notion of the bias and presenting an intuitive explanation why the sample mean is based in MABs, especially under adaptive sampling, stopping and choosing rules. Let $\widehat{\mu}_\kappa(\mathcal{T})$ be the sample mean of chosen arm κ at a stopping time $\mathcal{T}$. We first formally define the bias of the sample mean as follows:

Definition 1.1. *The (unconditional) bias of the sample mean of chosen arm κ at a stopping time $\mathcal{T}$ is defined by*

$$\mathbb{E}\left[\widehat{\mu}_\kappa(\mathcal{T}) - \mu_\kappa\right] = \sum_{k=1}^K \mathbb{P}(\kappa = k)\left\{\mathbb{E}\left[\widehat{\mu}_k(\mathcal{T}) \mid \kappa = k\right] - \mu_k\right\}, \tag{1.4}$$

where each $\mathbb{E}\left[\widehat{\mu}_k(\mathcal{T}) \mid \kappa = k\right] - \mu_k$ is the bias of the sample mean of arm k conditioned on the selection event $\{\kappa = k\}$. For general conditioning event C, the conditional bias of the sample mean of arm k defined by

$$\mathbb{E}\left[\widehat{\mu}_k(\mathcal{T}) \mid C\right] - \mu_k, \tag{1.5}$$

for each $k \in [K]$.

Note that the unconditional bias accounts for all possible outcomes of the MAB experiment. On the other hand, for the conditional bias case, the sample means of the arms are counted only when certain outcomes have occurred.

Biases induced by adaptive stopping and choosing rules are more straightforward to understand than the bias from the adaptive sampling rule since many MAB algorithms seem to cherry-pick a specific arm or time index at which the sample mean of the chosen arm at the chosen time looks like significantly different from the one of other arms at previous times. As an example of the bias induced by an adaptive stopping rule, consider sequential testing procedures for detecting the mean difference between two arms. In this case, most testing procedures are designed to be stopped and reject the null (no difference) whenever a test statistic based on sample means are crossing a boundary. Since the test statistic can cross boundary not only because of the underlying true mean difference but also because of random fluctuations, we can expect that the sample mean is based when the test stopped.

Similarly, as an example of the bias induced by an adaptive choosing rule, suppose we can collect the same number of observations from each arm and choose the one with the largest sample mean as the best arm. Again, since the largest sample mean can result both from sizes of true means of underlying arms and randomness of observations, the sample mean would be biased for the chosen arm.

The bias from the adaptive sampling is more subtle. Suppose we are interested in a predetermined arm k at a fixed time T. Then, the sample mean of arm k at time t can be written as follows:

$$\begin{aligned}\widehat{\mu}_k(T) &= \frac{1}{N_k(T)} \sum_{t=1}^{T} \mathbb{1}(A_t = k) Y_t \\ &= \frac{1}{N_k(T)} \sum_{i=1}^{N_k(T)} X_{i,k}^*,\end{aligned}$$

where Y_t is the observation at time t and $X_{i,k}^*$ is the i-th observation from arm k for $i = 1, \ldots, N_k(T)$. From the second expression of the sample mean, we know that the sample mean can be biased if each $X_{i,k}^*$ and $N_k(T)$ are correlated. That is, the adaptive sampling rule can induce the bias of the sample mean if the sampling rule decides whether to draw more or less samples depending on each current observation from arm k. For example, in many regret minimization MAB algorithms including ϵ-greedy, UCB and Thompson sampling, the sample rule is designed to draw more samples from arms whose sample means larger than the other. Hence, this type of sampling rule would induce a negative correlation between $1/N_k(T)$ and $X_{i,k}^*$ which can result in the negative bias of the sample mean.

These intuitive explanations why the sample mean is based in MABs will be formalized by introducing the notion of "optimality" and "monotonicity" properties of rules in the following chapter.

Chapter 2

Sign of the unconditional bias of the sample mean

In this chapter, we provide a comprehensive understanding of the sign of the (unconditional) bias of the sample mean, decoupling the effects of adaptive sampling, stopping and choosing. Specifically, in a general and unified MAB framework, we first define natural notions of monotonicity (a special case of which we call "optimism") of sampling, stopping and choosing rules. Under no assumptions on the distributions beyond assuming that their means exist, we show that optimistic sampling provably results in a negative bias, but optimistic stopping and optimistic choosing both provably result in a positive bias. Thus, the net bias can be positive or negative in general. This message is in contrast to a recent thought-provoking work by Nie et al. [2018] titled *"Why adaptively collected data has a negative bias..."* that is unfortunately misleading for practitioners, since it only analyzed the bias of adaptive sampling for a fixed arm at a fixed time.

As a concrete example, consider an offline analysis of data that was collected by an MAB algorithm (with any aim). Suppose that a practitioner wants to estimate the mean reward of some of the better arms that were picked more frequently by the algorithm. Nie et al. [2018] proved that the sample mean of each arm is negatively biased under fairly common adaptive sampling rules. Although this result is applicable only to a fixed arm at a fixed time, it could instill a possibly false sense of comfort with sample mean estimates since the practitioner might possibly think that sample means are underestimating the effect size. However, we prove that if the algorithm was adaptively stopped and the arm index was adaptively picked, then the net bias can actually be positive. Indeed, we prove that this is the case for the lil'UCB algorithm (Corollary 2.12), but it is likely true more generally as captured by our main theorem. Thus, the sample mean may actually overestimate the effect size. This is an important and general phenomenon for both theoreticians (to study further and quantify) and for practitioners (to pay heed to) because if a particular arm is later deployed in practice, it may yield a lower reward than was possibly expected from the offline analysis.

Related work and our contributions. Adaptive mean estimation, in each of the three senses described above, has received much attention in both recent and past literature. Below, we discuss how our work relates to past work, proceeding one notion at a time in approximate historical order.

We begin by noting that a single-armed bandit is simply a random walk, where adaptive stopping has been extensively studied. The book by Gut [2009] on stopped random walks is an excellent reference, summarizing almost 60 years of advances in sequential analysis. Most of these extensive results on random walks have not been extended to the MAB setting, which naturally involves adaptive sampling and choosing.

Of particular relevance is the paper by Starr and Woodroofe [1968] on the sign of the bias under adaptive stopping, whose work is subsumed by ours in two ways: we not only extend their insights to the MAB setting, but even for the one-armed setting, our results generalize theirs.

Characterizing the sign of the bias of the sample mean under adaptive sampling has been a recent topic of interest due to a surge in practical applications. While estimating MAB ad revenues, Xu et al. [2013] gave an informal argument of why the sample mean is *negatively* biased for "optimistic" algorithms. Later, Villar et al. [2015] encountered this negative bias in a simulation study motivated by using MAB for clinical trials. Most recently, Bowden and Trippa [2017] derived an exact formula for the bias and Nie et al. [2018] formally provided conditions under which the bias is negative. Our results on "optimistic" sampling inducing a negative bias generalize the corresponding results in these past works.

Most importantly, however, these past results hold only at a predetermined time and for a fixed arm. Here, we put forth a complementary viewpoint that "optimistic" stopping and choosing induces a *positive* bias. Indeed, one of our central conceptual contributions is an appropriate and crisp definition of "monotonicity" and "optimism" (Definition 2.1), that enables a clean and general analysis.

Our main theoretical result, Theorem 2.10, allows the determination of the sign of the bias in several interesting settings. Importantly, the bias may be of any sign when optimistic sampling, stopping and choosing are all employed together. We demonstrate the practical validity of our theory using some simulations that yield interesting insights in their own right.

2.1 The sign of the bias under adaptive sampling, stopping and choosing

2.1.1 Examples of positive bias due to "optimistic" stopping or choosing

In MAB problems, collecting higher rewards is a common objective of adaptive sampling rules, and hence they are often designed to sample more frequently from a distribution which has larger sample mean than the others. Nie et al. [2018] proved that the bias of the sample mean for any *fixed* arm and at any *fixed* time is negative when the sampling rule satisfies two conditions called "Exploit" and "Independence of Irrelevant Options" (IIO). However, the emphasis on *fixed* is important: their conditions are not enough to determine the sign of the bias under adaptive stopping or choosing, even in the simple nonadaptive sampling setting. Before formally defining our crucial notions of "optimism" in the next subsection, it is instructive to look at some examples.

Example 2.1. *Suppose we continuously alternate between drawing a sample from each of two Bernoulli distributions with mean parameters $\mu_1, \mu_2 \in (0,1)$. This sampling rule is fully deterministic, and thus it satisfies the Exploit and IIO conditions in Nie et al. [2018]. For any fixed time t, the bias equals zero for both sample means. Define a stopping time $\mathcal{T}$ as the first time we observe $+1$ from the first arm. Then the sample size of the first arm, $N_1(\mathcal{T})$, follows a geometric distribution with parameter μ_1, which implies that the bias of $\widehat{\mu}_1(\mathcal{T})$ is*

$$\mathbb{E}\left[\widehat{\mu}_1(\mathcal{T}) - \mu_1\right] = \mathbb{E}\left[\frac{1}{N_1(\mathcal{T})}\right] - \mu_1 = \frac{\mu_1 \log(1/\mu_1)}{1-\mu_1} - \mu_1,$$

which is positive for all $\mu_1 \in (0,1)$.

This example shows that for nonadaptive sampling, adaptive stopping can induce a *positive* bias. In fact, this example is not atypical, but is an instance of a more general phenomenon explored in the one-armed setting in sequential analysis. For example, Siegmund [1978, Ch. 3] contains the following classical result for a Brownian motion $W(t)$ with positive drift $\mu > 0$.

Example 2.2. *If we define a stopping time as the first time $W(t)$ exceeds a line with slope η and intercept $b > 0$, that is $\mathcal{T}_B := \inf\{t \geq 0 : W(t) \geq \eta t + b\}$, then for any slope $\eta \leq \mu$, we have $\mathbb{E}\left[\frac{W(\mathcal{T}_B)}{\mathcal{T}_B} - \mu\right] = 1/b$. Note that a sum of Gaussians with mean μ behaves like a time-discretization of a Brownian motion with drift μ; since $\mathbb{E}W(t) = t\mu$, we may interpret $W(\mathcal{T}_B)/\mathcal{T}_B$ as a stopped sample mean, and the last equation implies that its bias is $1/b$, which is positive.*

Generalizing further, Starr and Woodroofe [1968] proved the following remarkable result.

Example 2.3. *If we stop when the sample mean crosses any predetermined upper boundary, the stopped sample mean is always positive biased (whenever the stopping time is a.s. finite). Explicitly, choosing any arbitrary sequence of real-valued constants $\{c_k\}$, define $\mathcal{T}_c := \inf\{t : \widehat{\mu}_1(t) > c_t\}$, then as long as the observations X_i have a finite mean and $\mathcal{T}_c$ is a.s. finite, we have $\mathbb{E}\left[\widehat{\mu}_1(\mathcal{T}_c)\right] \geq \mu_1$.*

Surprisingly, we will generalize the above strong result even further. Additionally, stopping times in the MAB literature can be thought of as extensions of $\mathcal{T}_c$ and $\mathcal{T}_B$ to a setting with multiple arms, and we will prove that indeed the bias induced will still be positive. We end with an example of the positive bias induced by "optimistic" choosing:

Example 2.4. *Given K standard normals $\{Z_i\}$ (to be thought of as one sample from each of K arms), let $\kappa = \arg\max_k Z_k$, that is, we choose the arm with the largest observation. It is well known that $\mathbb{E}\left[Z_\kappa\right] = \mathbb{E}\left[\max_{k\in[K]} Z_k\right] \asymp \sqrt{2\log K}$. Since $\mathbb{E}Z_k = 0$ for all k, but $\mathbb{E}Z_\kappa > 0$, the "optimistic" choice κ induces a positive bias.*

In many typical MAB settings, we should expect sample means to have two contradictory sources of bias: negative bias from "optimistic sampling" and positive bias from "optimistic stopping/choosing".

2.1.2 Positive or negative bias under monotonic sampling, stopping and choosing

Based on the expression (1.3), we formally state a characteristic of MAB algorithms which fully determines the sign of the bias as follows.

Definition 2.1. *For each $k \in [K]$, we say a MAB algorithm satisfies the "monotonic increasing (or decreasing) property for arm k" if for any $i \in \mathbb{N}$, the function $\mathcal{D}_\infty^* \mapsto \mathbb{1}\left(\kappa = k\right)/N_k(\mathcal{T})$, is an increasing (or decreasing) function of $X_{i,k}^*$ while keeping all other entries in $\mathcal{D}_\infty^*$ fixed. Further, we say that*

- *A sampling rule is optimistic for arm k if the function $\mathcal{D}_\infty^* \mapsto N_k(t)$ is an increasing function of $X_{i,k}^*$ while keeping all other entries in $\mathcal{D}_\infty^*$ fixed for any fixed $i \in \mathbb{N}$ and $t \geq 1$;*
- *A stopping rule is optimistic for arm k if the function $\mathcal{D}_\infty^* \mapsto \mathcal{T}$ is a decreasing function of $X_{i,k}^*$ while keeping all other entries in $\mathcal{D}_\infty^*$ fixed for any fixed $i \in \mathbb{N}$;*
- *A choosing rule is optimistic for arm k if the function $\mathcal{D}_\infty^* \mapsto \mathbb{1}(\kappa = k)$ is an increasing function of $X_{i,k}^*$ while keeping all other entries in $\mathcal{D}_\infty^*$ fixed for any fixed $i \in \mathbb{N}$.*

Note that if a MAB algorithm consists of an optimistic sampling (or stopping or choosing) rule, with the other components being nonadaptive, then the algorithm satisfies is the monotonically decreasing (increasing) property. We remark that nonadaptive just means independent of the entries $X_{i,k}^*$, but it is not necessarily deterministic[1]. The above definition warrants some discussion to provide intuition.

Roughly speaking, with an optimistic stopping, if a sample from the k-th distribution was increased while keeping all other values fixed, the algorithm would reach its termination criterion sooner. For instance, $\mathcal{T}_B$ from Example 2.2 and the criterion in Example 2.1 are both optimistic stopping rules. Most importantly, boundary-crossing is optimistic:

[1] An example of a random but nonadaptive stopping rule: flip a (potentially biased) coin at each step to decide whether to stop. An example of a random but nonadaptive sampling rule: with probability half pick a uniformly random arm, and with probability half pick the arm that has been sampled most often thus far.

Fact 2.5. *The general boundary-crossing stopping rule of Starr and Woodroofe [1968], denoted $\mathcal{T}_c$ in Example 2.3, is an optimistic stopping rule (and hence optimistic stopping is a weaker condition).*

Optimistic stopping rules do not need to be based on the sample mean; for example, if $\{c_t\}$ is an arbitrary sequence, then $\mathcal{T} := \inf\{t \geq 3 : X_t + X_{t-2} \geq c_t\}$ is an optimistic stopping rule. In fact, $\mathcal{T}_\ell := \inf\{t \geq 3 : \ell_t(X_1, \ldots, X_t) \geq c_t\}$ is optimistic, as long as each ℓ_t is coordinatewise nondecreasing.

For optimistic choosing, the previously discussed argmax rule (Example 2.4) is optimistic. More generally, it is easy to verify the following:

Fact 2.6. *For any probabilities $p_1 \geq p_2 \cdots \geq p_K$ that sum to one, a rule that chooses the arm with the k-th largest empirical mean with probability p_k, is an optimistic choosing rule.*

Turning to the intuition for optimistic sampling, if a sample from the k-th distribution was increased while keeping all other values fixed, the algorithm would sample the k-th arm more often. We claim that optimistic sampling is a weaker condition than the Exploit and IIO conditions employed by Nie et al. [2018].

Fact 2.7. *The "Exploit" and "IIO" conditions in Nie et al. [2018] together imply that the sampling rule is optimistic (and hence optimistic sampling is a weaker condition). Further, as summarized in Appendix A.1, ϵ-greedy, UCB and Thompson sampling (Gaussian-Gaussian and Beta-Bernoulli, for instance) are all optimistic sampling rules.*

For completeness, we prove the first part formally in Appendix A.1.1, which builds heavily on observations already made in the proof of Theorem 1 in Nie et al. [2018]. Beyond the instances mentioned above, Corollary A.2 in the supplement captures a sufficient condition for Thompson sampling with one-dimensional exponential families and conjugate priors to be optimistic. We now provide an expression for the bias that holds at any stopping time and for any sampling rule.

Proposition 2.8. *Let $\mathcal{T}$ be a stopping time with respect to the natural filtration $\{\mathcal{F}_t\}$. For each fixed $k \in [K]$ such that $0 < \mathbb{E}N_k(\mathcal{T}) < \infty$, the bias of $\widehat{\mu}_k(\mathcal{T})$ is given as*

$$\mathbb{E}\left[\widehat{\mu}_k(\mathcal{T}) - \mu_k\right] = -\frac{\mathrm{Cov}\left(\widehat{\mu}_k(\mathcal{T}), N_k(\mathcal{T})\right)}{\mathbb{E}\left[N_k(\mathcal{T})\right]}. \tag{2.1}$$

The proof may be found in Appendix A.2.3. A similar expression was derived in Bowden and Trippa [2017], but only for a fixed time T. In order to extend it to stopping times (that are allowed to be infinite, as long as $\mathbb{E}N_k(\mathcal{T}) < \infty$), we derive a simple generalization of Wald's first identity to the MAB setting. Specifically, recalling that $S_k(t) = \widehat{\mu}_k(t)N_k(t)$, we show the following:

Lemma 2.9. *Let $\mathcal{T}$ be a stopping time with respect to the natural filtration $\{\mathcal{F}_t\}$. For each fixed $k \in [K]$ such that $\mathbb{E}N_k(\mathcal{T}) < \infty$, we have $\mathbb{E}[S_k(\mathcal{T})] = \mu_k\mathbb{E}[N_k(\mathcal{T})]$.*

This lemma is also proved in Appendix A.2.3. Proposition 2.8 provides a simple, and somewhat intuitive, expression of the bias for each arm. It implies that if the covariance of the sample mean of an arm and the number of times it was sampled is positive (negative), then the bias is negative (positive). We now formalize this intuition below, including for adaptively chosen arms. The following theorem shows that if a MAB algorithm satisfies the monotonically increasing (or decreasing) property then the sample mean is positively (or negatively) biased.

Theorem 2.10. *Let $\mathcal{T}$ be a stopping time with respect to the natural filtration $\{\mathcal{F}_t\}$ and let $\kappa : \mathcal{D}_\mathcal{T} \mapsto [K]$ be a choosing rule. Suppose each arm has finite expectation and, for all k with $\mathbb{P}(\kappa = k) > 0$, we have $\mathbb{E}\left[N_k(\mathcal{T})\right] < \infty$ and $N_k(\mathcal{T}) \geq 1$. If, for each arm, the MAB algorithm satisfies the monotonically decreasing property, for example under optimistic sampling with nonadaptive stopping and choosing, then we have*

$$\mathbb{E}\left[\widehat{\mu}_\kappa(\mathcal{T}) \mid \kappa = k\right] \leq \mu_k, \quad \forall k : \mathbb{P}(\kappa = k) > 0, \tag{2.2}$$

which also implies that

$$\mathbb{E}\left[\widehat{\mu}_\kappa(\mathcal{T}) - \mu_\kappa\right] \leq 0. \tag{2.3}$$

Similarly if, for each arm, the MAB algorithm satisfies the monotonically increasing property, for example under optimistic stopping with nonadaptive sampling and choosing, or under optimistic choosing with nonadaptive sampling and stopping, then we have

$$\mathbb{E}\left[\widehat{\mu}_\kappa(\mathcal{T}) \mid \kappa = k\right] \geq \mu_k, \ \ \forall k : \mathbb{P}(\kappa = k) > 0, \tag{2.4}$$

which also implies that

$$\mathbb{E}\left[\widehat{\mu}_\kappa(\mathcal{T}) - \mu_\kappa\right] \geq 0. \tag{2.5}$$

If each arm has a bounded distribution then the condition $\mathbb{E}\left[N_k(\mathcal{T})\right] < \infty$ can be dropped.

Remark 2.11. *In fact, if each arm has a finite p-th moment for a fixed $p > 2$ then the condition $\mathbb{E}\left[N_k(\mathcal{T})\right] < \infty$ can be dropped.*

The proofs of Theorem 2.10 and Remark 2.11 can be found in Appendix A.2.1. See also Appendix A.1.3 for an intuitive explanation of the sign of the bias under optimistic sampling, stopping or choosing rules. The expression (2.1) intuitively suggests situations when the sample mean estimator $\widehat{\mu}_k(\mathcal{T})$ is biased, while the inequalities in (2.2) and (2.4) determine the direction of bias under the monotonic or optimistic conditions. Due to Facts 2.5, 2.6 and 2.7, several existing results are immediately subsumed and generalized by Theorem 2.10. Further, the following corollary is a particularly interesting special case dealing with the lil'UCB algorithm by Jamieson et al. [2014] which consists of adaptive sampling, stopping and choosing rules, as summarized in Section 2.2.3.

Corollary 2.12. *The lil'UCB algorithm satisfies the monotonically increasing property, and thus the sample mean of the reported arm when lil'UCB stops is always positively biased.*

The proof is described in Appendix A.2.2. The above result is interesting because of the following reasons: (a) when viewed separately, the sampling, stopping and choosing rules of the lil'UCB algorithm all are optimistic; hence it is apriori unclear which rule dominates and whether the net bias should be positive or negative; (b) we did not have to alter anything about the algorithm in order to prove that it is a monotonically increasing strategy (for any distribution over arms, for any number of arms). The generality of the above result showcases the practical utility of our theorem, whose message is in sharp contrast to the title of the paper by Nie et al. [2018].

Next, we provide simulation results that verify that our monotonic and optimistic conditions accurately capture the sign of the bias of the sample mean.

2.2 Numerical experiments

2.2.1 Negative bias from optimistic sampling rules in multi-armed bandits

Recall Fact 2.7, which stated that common MAB adaptive sampling rules like greedy (or ϵ-greedy), upper confidence bound (UCB) and Thompson sampling are optimistic. Thus, for a deterministic stopping time, Theorem 2.10 implies that the sample mean of each arm is always negatively biased. To demonstrate this, we conduct a simulation study in which we have three unit-variance Gaussian arms with $\mu_1 = 1, \mu_2 = 2$ and $\mu_3 = 3$. After sampling once from each arm, greedy, UCB and Thompson sampling are used to continue sampling until $T = 200$. We repeat the whole process from scratch 10^4 times for each algorithm to get an accurate estimate for the bias.[2] Due to limited space, we present results from UCB and Thompson

[2] In all experiments, sizes of reported biases are larger than at least 3 times the Monte Carlo standard error.

sampling only but detailed configurations of algorithms and a similar result for the greedy algorithm can be found in Appendix A.3.1. Figure 2.1 shows the distribution of observed differences between sample means and the true mean for each arm. Vertical lines correspond to biases. The example demonstrates that the sample mean is negatively biased under optimistic sampling rules.

Remark 2.13. *The main goal in our simulations is to visualize and corroborate our theoretical results about the sign of the bias. As a result, we do not make any attempt to optimize the parameters for UCB or Thompon sampling for the purpose of minimizing the regret, since the latter is not this chapter's aim. However, investigating the relationship between the performance of MAB algorithms and the bias at the time horizon would be an interesting future direction of research.*

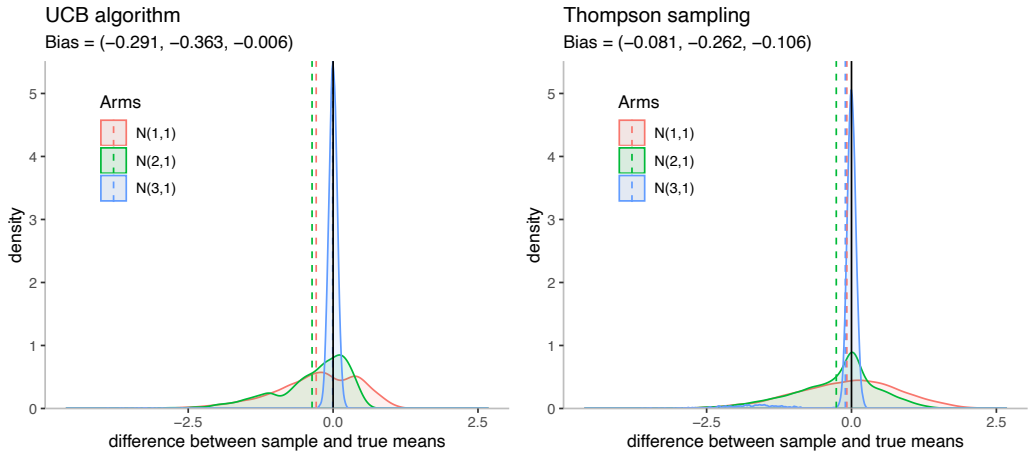

Figure 2.1: *Data is collected by UCB (left) and Thompson sampling (right) algorithms from three unit-variance Gaussian arms with $\mu_1 = 1, \mu_2 = 2$ and $\mu_3 = 3$. For all three arms, sample means are negatively biased (at fixed times). A similar result for the greedy algorithm can be found in Appendix A.3.1.*

2.2.2 Bias from stopping a one-sided sequential likelihood ratio test

Suppose we have two independent sub-Gaussian arms with common and known parameter σ^2 but unknown means μ_1 and μ_2. Consider the following testing problem:

$$H_0 : \mu_1 \leq \mu_2 \;\text{ vs }\; H_1 : \mu_1 > \mu_2.$$

To test this hypothesis, suppose we draw a sample from arm 1 for every odd time and from arm 2 for every even time. Instead of conducting a test at a fixed time, we can use the following one-sided sequential likelihood ratio test [Robbins, 1970, Howard et al., 2018b]: for any fixed $w > 0$ and $\alpha \in (0, 1)$, define a stopping time $\mathcal{T}$ as

$$\mathcal{T}^w := \inf\left\{ t \in \mathbb{N}_{\text{even}} : \widehat{\mu}_1(t) - \widehat{\mu}_2(t) \geq \frac{2\sigma}{t}\sqrt{(t+2w)\log\left(\frac{1}{2\alpha}\sqrt{\frac{t+2w}{2w}}+1\right)}\right\}, \tag{2.6}$$

where $\mathbb{N}_{\text{even}} := \{2n : n \in \mathbb{N}\}$. For a given fixed maximum even time $M \geq 2$, we stop sampling at time $\mathcal{T}^w_M := \min\{\mathcal{T}^w, M\}$. Then, we reject the null H_0 if $\mathcal{T}^w_M < M$. It can be checked [Howard et al., 2018b, Section 8] that, for any fixed $w > 0$, this test controls the type-1 error at level α and the power goes to 1 as M goes to infinity.

For the arms 1 and 2, these are special cases of optimistic and pessimistic stopping rules respectively. From Theorem 2.10, we have that $\mu_1 \leq \mathbb{E}\widehat{\mu}_1(\mathcal{T}_M^w)$ and $\mu_2 \geq \mathbb{E}\widehat{\mu}_2(\mathcal{T}_M^w)$. To demonstrate this, we conduct two simulation studies with unit variance Gaussian errors: one under the null hypothesis $(\mu_1, \mu_2) = (0, 0)$, and one under the alternative hypothesis $(\mu_1, \mu_2) = (1, 0)$. We choose $M = 200$, $w = 10$ and $\alpha = 0.1$. As before, we repeat each experiment 10^4 times for each setting. Figure 2.2 shows the distribution of observed differences between sample means and the true mean for each arm under null and alternative hypothesis cases. Vertical lines correspond to biases. The simulation study demonstrates that the sample mean for arm 1 is positively biased and the sample mean for arm 2 is negatively biased as predicted.

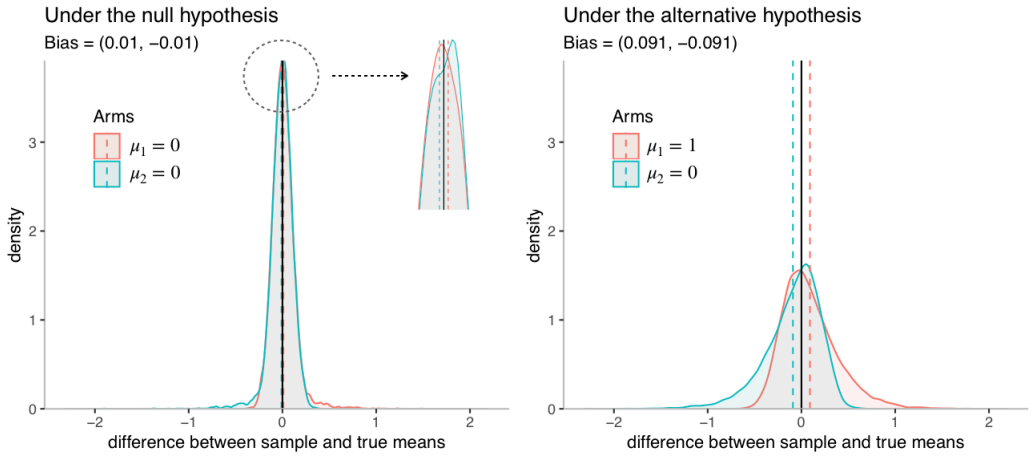

Figure 2.2: *Data is collected from the one-sided sequential likelihood ratio test procedure described in Section 2.2.2. The sample mean for arm 1 is positively biased and the sample mean for arm 2 is negatively biased under both null and alternative hypothesis cases. Note that the size of the bias under the null hypothesis is smaller than the one under the alternative hypothesis since the number of collected samples is larger under the null hypothesis.*

2.2.3 Positive bias of the lil'UCB algorithm in best-arm identification

Suppose we have K sub-Gaussian arms with mean $\mu_1, \ldots, \mu_K$ and known parameter σ. In the best-arm identification problem, our target of inference is the arm with the largest mean. There exist many algorithms for this task including lil'UCB [Jamieson et al., 2014], Top-Two Thompson Sampling [Russo, 2016] and Track-and-Stop [Garivier and Kaufmann, 2016].

In Corollary 2.12, we showed that the lil'UCB algorithm is monotonically increasing, and thus the sample mean of the chosen arm is positively biased. In this subsection, we verify it with a simulation. It is an interesting open question whether different types of best-arm identification algorithms also yield positively biased sample means.

The lil'UCB algorithm consists of the following optimistic sampling, stopping and choosing rules:

- Sampling: For any $k \in [K]$ and $t = 1, \ldots K$, define $\nu_t(k) = \mathbb{1}(t = k)$. For $t > K$,

$$\nu_t(k) = \begin{cases} 1 & \text{if } k = \arg\max_{j \in [K]} \widehat{\mu}_j(t-1) + u_t^{\text{lil}}\left(N_j(t-1)\right), \\ 0 & \text{otherwise,} \end{cases}$$

  where $\delta, \epsilon, \lambda$ and β are algorithm parameters and

$$u_t^{\text{lil}}(n) := (1+\beta)(1+\sqrt{\epsilon})\sqrt{2\sigma^2(1+\epsilon)\log\left(\log((1+\epsilon)n)/\delta\right)/n}.$$

- Stopping: $\mathcal{T} = \inf\left\{t > K : N_k(t) \geq 1 + \lambda \sum_{j \neq k} N_j(t) \text{ for some } k \in [K]\right\}$.
- Choosing: $\kappa = \arg\max_{k \in [K]} N_k(\mathcal{T})$.

Once we stop sampling at time $\mathcal{T}$, the lil'UCB algorithm guarantees that κ is the index of the arm with largest mean with some probability depending on input parameters. Based on this, we can also estimate the largest mean by the chosen stopped sample mean $\widehat{\mu}_\kappa(\mathcal{T})$. The performance of this sequential procedure can vary based on underlying distribution of the arm and the choice of parameters. However, we can check this optimistic sampling and optimistic stopping/choosing rules which would yield negative and positive biases respectively, jointly satisfy the monotonically increasing property and thus the chosen stopped sample mean $\widehat{\mu}_\kappa(\mathcal{T})$ is always positively biased for any choice of parameters.

To verify it with a simulation, we set 3 unit-variance Gaussian arms with means $(\mu_1, \mu_2, \mu_3) = (g, 0, -g)$ for each gap parameter $g = 1, 3, 5$. We conduct 10^4 trials of the lil'UCB algorithm with a valid choice of parameters described in Jamieson et al. [2014, Section 5]. Figure 2.3 shows the distribution of observed differences between the chosen sample means and the corresponding true mean for each δ. Vertical lines correspond to biases. The simulation study demonstrates that, in all configurations, the chosen stopped sample mean $\widehat{\mu}_\kappa(\mathcal{T})$ is always positively biased. (see Appendix A.2.2 for a formal proof.)

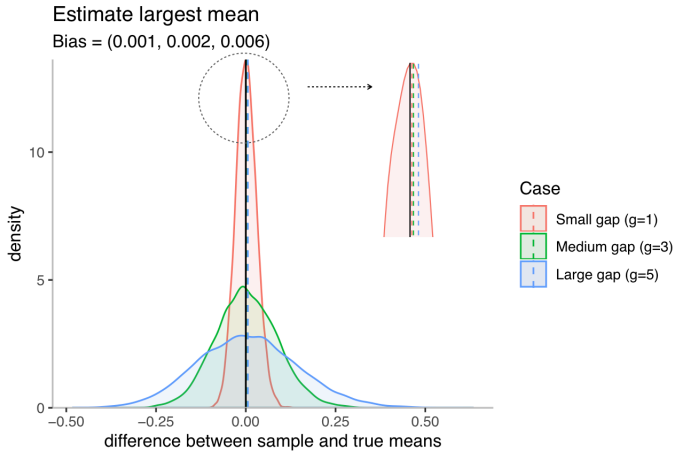

Figure 2.3: *Data is collected by the lil'UCB algorithm run on three unit-variance Gaussian arms with $\mu_1 = g, \mu_2 = 0$ and $\mu_3 = -g$ for each gap parameter $g = 1, 3, 5$. For all cases, chosen sample means are positively biased. The bias is larger for a larger gap since the number of collected samples is smaller on an easier task.*

2.3 Summary

In this chapter, we provide a general and comprehensive characterization of the sign of the bias of the sample mean in multi-armed bandits. Our main conceptual innovation was to define new weaker conditions (monotonicity and optimism) that capture a wide variety of practical settings in both the random walk (one-armed bandit) setting and the MAB setting. Using this, our main theoretical contribution, Theorem 2.10, significantly generalizes the kinds of algorithms or rules for which we can mathematically determine the sign of the bias for any problem instance. Our simulations confirm the accuracy of our theoretical predictions for a variety of practical situations for which such sign characterizations were previously unknown.

Chapter 3

Sign of the conditional bias of sample mean and CDF

In the previous chapter, we studied the bias of the sample mean and showed that the sign of the bias can be related to adaptive sampling, stopping and choosing in precise ways. However, the previous theoretical understanding of the bias in MAB experiments is limited in two aspects. First, virtually all results concern the bias in mean estimation and do not cover other functionals of the arms. Secondly, and perhaps more interestingly, the existing guarantees cover only the *unconditional bias* of the sample mean, i.e., the bias obtained by accounting for all possible outcomes of the MAB experiment. However, in practice, one is often interested in the sample means of the arms only when certain outcomes have occurred. For instance, the analyst may wish to evaluate the sample mean of a given arm only when that arm was identified as the best arm or, in a sequential framework corresponding to a MAB experiment with only one arm, when the null hypothesis has been rejected or when the random criterion for determining whether enough samples have been collected has been met. In all these cases, it is of interest to compute the *conditional bias*, i.e., the bias of the sample mean given a certain *conditioning event,* such as that the arm of interest turned out to be the best arm. A priori, it is not at all clear how the sign of the conditional bias is affected by the choice of the conditioning event and by other sources of data adaptivity (e.g., sampling and stopping), or whether the signs of the conditional and unconditional bias should match.

As a concrete example, suppose we have K prototypes of an online service and wish to test whether the potential average revenue of each prototype (i.e., arm), μ_k, would be larger than a pre-specified threshold $\mu_0 > 0$ based on a stream of test user samples. The usual sequential testing approach will be based on an appropriate upper stopping boundary and will reject each null $\mu_k \leq \mu_0$ if the corresponding sample mean has crossed such boundary during the testing period. If the null is rejected, then we conclude the prototype as a promising one. If not, we disregard it. It is well known [see, e.g., Starr and Woodroofe, 1968, Shin et al., 2019a] that for each prototype, the sample mean based on data collected through this sequential testing procedure is positively biased: that is, for each $k \in [K] := \{1, \ldots, K\}$, $\mathbb{E}\left[\widehat{\mu}_k\right] \geq \mu_k$ regardless of whether the true mean μ_k is larger or smaller than the threshold μ_0. This positive bias result can provide a useful warning signal about possible overestimation of the true revenues. At the same time, however, without careful consideration of the conditioning effect, we may end up with a false sense of comfort with low sample mean estimates from "disregarded" prototypes. Indeed, we can naively expect the true revenues of the disregarded prototypes should be even lower than the observed estimates based on the positive bias result $\mathbb{E}\left[\widehat{\mu}_k\right] \geq \mu_k$. In fact, this would be a wrong conclusion: conditioned on the disregarding event C, the sample mean is negatively biased and we have $\mathbb{E}\left[\widehat{\mu}_k \mid C\right] \leq \mu_k$, as demonstrated in Section 3.2.1.

In this chapter, we make the following contributions:

- We derive in Theorem 3.1 sufficient conditions for determining the sign of the conditional bias that hold under arbitrary conditioning events and rely on certain natural, highly interpretable monotonicity properties of the rules used for adaptive sampling, stopping and selecting. Our analysis captures in a rigorous and intuitive way the interaction between adaptivity arising from the data collection stage (sampling and stopping) with adaptivity in the choice of the target for inference.
- We characterize the sign of the conditional bias of monotone functions of the rewards of each arm, which includes the sample mean as a special case.
- In Section 3.2, we demonstrate with several examples in best arm identification and sequential testing problems how the conditional and unconditional bias of the sample means of the arms can have opposite signs. These are, we believe, instances of a general, important phenomenon of theoretical and practical relevance.

Overall, our results advance our ability to assess the impact of the bias in adaptive data analysis problems and offer several new and interesting insights on this important issue.

3.1 The sign of the conditional bias

In this section, we generalize Theorem 2.10 and derive sufficient conditions for the sign of the conditional bias under arbitrary conditioning events of monotone functions of the arms rewards.

We first introduce some notation. For each arm k, let F_k denote the corresponding cumulative distribution function (CDF): $y \in \mathbb{R} \mapsto F_k(y) = \mathbb{P}(X \leq y)$, where $X \sim P_k$. Let $\widehat{F}_{k,t}$ be the empirical CDF for arm k based on samples up to time t; that is, $\widehat{F}_{k,t}$ is a random function on $\mathbb{R}$ and taking values in $[0,1]$ given by

$$y \in \mathbb{R} \mapsto \widehat{F}_{k,t}(y) := \frac{1}{N_k(t)} \sum_{s=1}^{t} \mathbb{1}\left(A_s = k, Y_s \leq y\right), \tag{3.1}$$

which returns the fractions of samples from arm k whose values are no larger than y. For a stopping time $\mathcal{T}$ with respect to the MAB filtration $\{\mathcal{F}_t\}$, we then define $\widehat{F}_{k,\mathcal{T}}$ to be the empirical CDF of the rewards of arm k up to time $\mathcal{T}$; that is, for each $y \in \mathbb{R}$,

$$\widehat{F}_{k,\mathcal{T}}(y) := \lim_{t\to\infty} \widehat{F}_{k,\mathcal{T}\wedge t}(y).$$

The use of the limit in the above definition is a necessary technicality in order to allow for the possibility that $\mathcal{T} = \infty$. Note that, on the event $\{N_k(\mathcal{T}) = \infty\} = \{\lim_{t\to\infty} N_k(\mathcal{T}\wedge t) = \infty\}$, we have that, for each $y \in \mathbb{R}$, $\widehat{F}_k(y) = F_k(y)$ almost surely, based on the strong law of large numbers; see Theorem 2.1 in Gut [2009].

Next, for any function $f : \mathbb{R} \to \mathbb{R}$ integrable with respect to P_k, let $E_k f = \int f(x)\mathrm{d}P_k(x)$ be the corresponding expectation. Similarly, we let

$$\widehat{E}_{k,t} f = \frac{1}{N_k(t)} \sum_{s=1}^{t} f(Y_s)\mathbb{1}\left(A_s = k\right)$$

to be the expectation of f under the empirical measure of the k-th arm at time t. Clearly, $\widehat{E}_{k,t} f$ is a random variable. If f is the identity function, then it is immediate to see that $E_k f = \mu_k$ and $\widehat{E}_{k,t} f = \widehat{\mu}_k(t)$. Also, for any $y \in \mathbb{R}$, setting f to be the indicator function $x \mapsto f(x) = \mathbb{1}(x \leq y)$ yields that $E_k f = F_k(y)$ and $\widehat{E}_{k,t} f = \widehat{F}_{k,t}(y)$. Finally, for a (possibly infinite) stopping time $\mathcal{T}$ with respect to the filtration $\{\mathcal{F}_t\}$, we set

$$\widehat{E}_{k,\mathcal{T}} f = \lim_{t\to\infty} \widehat{E}_{k,\mathcal{T}\wedge t} f$$

We are now ready to present the main result of this chapter. The proof is given in Appendix B.1.1. For any event C, we let $\mathbb{E}[\cdot|C]$ denote the conditional expectation given C.

Theorem 3.1. *Let $\mathcal{T}$ be a stopping time with respect to the MAB experiment natural filtration $\{\mathcal{F}_t\}$. For a fixed $k \in [K]$, suppose $N_k(\mathcal{T}) \geq 1$. Let $C \in \mathcal{F}_\mathcal{T}$ be any event at stopping time $\mathcal{T}$ with $\mathbb{P}(C) > 0$. Assume that, for each i, the function $\mathcal{D}_\infty^* \mapsto \mathbb{1}(C)/N_k(\mathcal{T})$ is a decreasing function of $X_{i,k}^*$ while keeping all other entries in $\mathcal{D}_\infty^*$ fixed. Then, for each $y \in \mathbb{R}$,*

$$\inf_{y \in \mathbb{R}} \left(\mathbb{E}\left[\widehat{F}_{k,\mathcal{T}}(y) \mid C \right] - F_k(y) \right) \geq 0, \tag{3.2}$$

or, equivalently, for any non-decreasing function $f : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}\left[\widehat{E}_{k,\mathcal{T}}|f| \mid C \right] < \infty$ and $E_{k,\mathcal{T}}|f| < \infty$,

$$\mathbb{E}\left[\widehat{E}_{k,\mathcal{T}} f \mid C \right] \leq E_k f. \tag{3.3}$$

Similarly, if, for each i, the function $\mathcal{D}_\infty^* \mapsto \mathbb{1}(C)/N_k(\mathcal{T})$ is an increasing function of $X_{i,k}^*$ while keeping all other entries in $\mathcal{D}_\infty^*$ fixed, then we have

$$\sup_{y \in \mathbb{R}} \left(\mathbb{E}\left[\widehat{F}_{k,\mathcal{T}}(y) \mid C \right] - F_k(y) \right) \leq 0, \tag{3.4}$$

or, equivalently,

$$\mathbb{E}\left[\widehat{E}_{k,\mathcal{T}} f \mid C \right] \geq E_k f, \tag{3.5}$$

for any non-decreasing function $f : \mathbb{R} \to \mathbb{R}$ satisfying $\mathbb{E}\left[\widehat{E}_{k,\mathcal{T}}|f| \mid C \right] < \infty$ and $E_{k,\mathcal{T}}|f| < \infty$.

The important conclusion from the above theorem is that the conditional expected value of the empirical CDF of the rewards of any given arm can be stochastically smaller or greater than the true CDF of the corresponding arm. And furthermore, the sign of the bias can be determined in the basis of natural and often verifiable (as shown below) monotonicity conditions that depend on (i) the specific sampling and stopping rules deployed in the MAB experiment and (ii) the choice of the conditioning event C.

Note that if in the theorem we choose C as the conditioning event $\{\kappa = k\}$ and the function f as the identity function, Theorem 3.1 yields Theorem 2.10 as a special case (indeed the statements about the unconditional bias in that result follows from those on the conditional bias given the conditioning event $\{\kappa = k\}$). We emphasize that Theorem 3.1 is a strict generalization of Theorem 2.10 in at least two ways: (i) it provides a more general guarantee about the conditional bias of monotone functions of the empirical CDF of the arms as opposed to just the sample mean and (ii) it allows for virtually any conditioning event that depends on the outcome of the MAB experiment (formally, that is measurable with respect to $\mathcal{F}_\mathcal{T}$).

The monotonicity assumption in Theorem 3.1 captures in a mathematically concise yet intuitive manner how adaptivity in the data collection process combines with adaptivity in the selective data analysis, exemplified by the conditioning event, to affect the sign of the bias. The interaction between these two sources of adaptivity is, at least to us, not at all apparent and, in fact, rather subtle. To illustrate this phenomenon, in the next section, we apply Theorem 3.1 in several, fairly routine, situations in adaptive data analysis to demonstrates how conditional and unconditional biases may very well have opposite signs.

3.2 Applications

In this section, we discuss several practical examples of the conditional bias results in Theorem 3.1.

3.2.1 Conditional versus unconditional bias of a stopped sequential test

Suppose we have a stream of samples from a single arm with a finite mean μ. For each t, let $\widehat{\mu}(t)$ be the sample mean of the rewards observed up to time t.

To test whether μ is larger than a reference value μ_0, we may construct an upper boundary $t \mapsto U(t)$ and conclude that $\mu \geq \mu_0$ if the sample mean ever crosses the upper boundary.

Let $\mathcal{T}$ be the first time the sample mean crosses the boundary, i.e. $\mathcal{T} := \inf\{t \geq 1 : \widehat{\mu}(t) \geq U(t)\}$. The stopping time $\mathcal{T}$ is an example of an optimistic stopping and, from Theorem 2.10, we can check that the stopped sample mean $\widehat{\mu}(\mathcal{T})$ is always positively biased.

However, the stopping time $\mathcal{T}$ can be large or even infinite with non-zero probability. Thus, in practice, we may want to allow for the possibility of stopping the sequential test before reaching the stopping time. Let M be any fixed predefined time at which we stop the testing procedure (if still ongoing), and let $\mathcal{T}_M := \min\{\mathcal{T}, M\}$ be the corresponding stopping time which takes account the early stopping option. Again, by Theorem 2.10, we know that the stopped sample mean $\widehat{\mu}(\mathcal{T}_M)$ is still positively biased, i.e. $\mathbb{E}[\widehat{\mu}(\mathcal{T}_M)] \geq \mu$, since the function $\mathcal{D}_\infty^* \mapsto 1/\mathcal{T}_M$ is an increasing function of X_i^* while keeping all other entries in $\mathcal{D}_\infty^*$ fixed. From Theorem 3.1, we can also check that the expected empirical CDF at stopping time $\mathcal{T}_M$ is negatively biased:

$$\sup_{y \in \mathbb{R}} \left(\mathbb{E}\left[\widehat{F}_{k,\mathcal{T}_M}(y) \right] - F_k(y) \right) \leq 0. \tag{3.6}$$

Conditioned on the early stopping event $\{M < \mathcal{T}\}$, however, Theorem 3.1 shows that the early stopped sample mean and the empirical CDF are negatively and positively biased, respectively, since the function $\mathcal{D}_\infty^* \mapsto \frac{\mathbb{1}(M<\mathcal{T})}{\mathcal{T}_M} = \frac{\mathbb{1}(M<\mathcal{T})}{M}$ is an decreasing function of X_i^* while keeping all other entries in $\mathcal{D}_\infty^*$ fixed. That is,

$$\mathbb{E}[\widehat{\mu}(M) \mid M < \mathcal{T}] \leq \mu, \tag{3.7}$$

$$\inf_{y \in \mathbb{R}} \left(\mathbb{E}\left[\widehat{F}_{k,M}(y) \mid M < \mathcal{T} \right] - F_k(y) \right) \geq 0. \tag{3.8}$$

However, we can also check that the function $\mathcal{D}_\infty^* \mapsto \frac{\mathbb{1}(M \geq \mathcal{T})}{\mathcal{T}_M} = \frac{\mathbb{1}(M \geq \mathcal{T})}{\mathcal{T}}$ is an increasing function of X_i^*, which implies that, on the line-crossing event $\{M \geq \mathcal{T}\}$, the sample mean and empirical CDF are positively and negatively biased, respectively. Thus, depending on the conditioning event, the conditional bias can be positive or negative. Also note that, without the early stopping condition, $\widehat{\mu}(M)$ is an unbiased estimator of μ.

Experiment: We verify these facts with simulations, where we repeat a stopped sequential test 10^5 times. In each test, the arm corresponds to a standard normal distribution, and the upper boundary is constructed by the point-wise upper confidence bounds $t \mapsto U(t) = \frac{z_\alpha}{\sqrt{t}}$, where z_α is the α-upper quantile of the standard normal distribution. Note that this upper boundary does not yield a valid testing procedure since it inflates the type 1 error. However, we choose this boundary with unusual parameters $\alpha = 0.2$ and $M = 10$ to manifest the difference between conditional and unconditional biases.

Figure 3.1 shows the point-wise averages of the observed conditional and unconditional CDFs from the simulated 10^5 stopped sequential tests. The black dashed line refers to the true CDF of the underlying arm. The red line corresponds to the average of the empirical CDFs, which lies below the true CDF, as expected, and thus the corresponding mean bias is positive. The green and blue lines correspond to the averages of empirical CDFs conditioned on the early stopping and line-crossing events, respectively. As predicted, conditioned on the early stopping event, the empirical CDF and the sample mean are positively and negatively biased, respectively. In contrast, conditioned on the line-crossing event, the empirical CDF and the sample mean are negatively and positively biased, respectively.

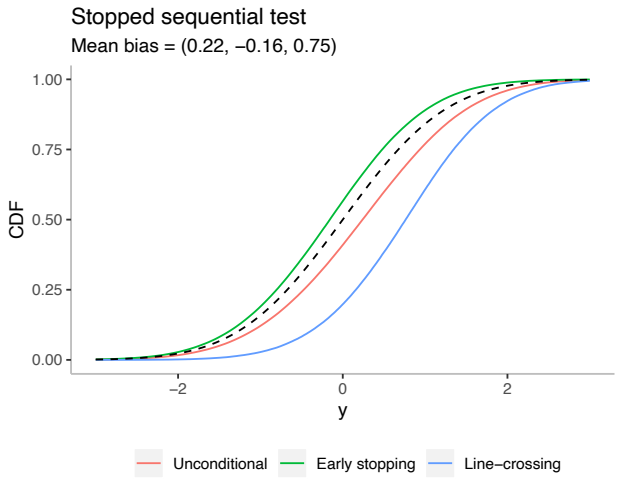

Figure 3.1: *Average of unconditional (red) and conditional (green and blue) empirical CDFs from repeated stopped sequential test as described in Section 3.2.1.*

3.2.2 Sequential test for two arms: conditional biases from upper and lower stopping boundaries

Suppose we have two independent arms with unknown means μ_1 and μ_2. In this subsection, we consider the following testing problem:

$$H_0 : \mu_1 \leq \mu_2 \;\text{ vs }\; H_1 : \mu_1 > \mu_2. \tag{3.9}$$

To test this hypothesis, we draw a sample from arm 1 for every odd time and from arm 2 for every even time. Then, at each even time t, we check whether the difference between the sample means $\widehat{\mu}_1(t)$, $\widehat{\mu}_2(t)$ from the two arms crosses predefined upper and lower stopping boundaries, $t \mapsto U(t)$ and $t \mapsto L(t)$.

To be specific, define stopping times $\mathcal{T}^U$ and $\mathcal{T}^L$ as follows:

$$\mathcal{T}^U := \inf\left\{t \in \mathbb{N}_{\text{even}} : \widehat{\mu}_1(t) - \widehat{\mu}_2(t) \geq U(t)\right\}, \tag{3.10}$$

$$\mathcal{T}^L := \inf\left\{t \in \mathbb{N}_{\text{even}} : \widehat{\mu}_1(t) - \widehat{\mu}_2(t) \leq L(t)\right\}. \tag{3.11}$$

Let $M > 0$ be a predetermined maximum time budget. Based on $\mathcal{T}^U, \mathcal{T}^L$ and M, we stop sampling whenever $\widehat{\mu}_1(t) - \widehat{\mu}_2(t)$ crosses one of the boundaries or the maximum time budget M is met. Define the corresponding stopping time as $\mathcal{T}_M := \min\left\{\mathcal{T}^U, \mathcal{T}^L, M\right\}$. At time $\mathcal{T}_M$, we accept H_1 if $\mathcal{T}_M = \mathcal{T}^U$ (upper-crossing event), and accept H_0 if $\mathcal{T}_M = \mathcal{T}^L$ (lower-crossing event). Otherwise, we declare that we do not have enough evidence to accept either one of two hypotheses.

In this case, we cannot apply Theorem 2.10 to determine the sign of the unconditional bias since the stopping rule is neither optimistic nor pessimistic, unconditionally. However, we can determine the sign of the conditional bias based on Theorem 3.1 since the the function $\mathcal{D}_\infty^* \mapsto \frac{\mathbb{1}(\mathcal{T}^U \leq \min\left\{\mathcal{T}^L, M\right\})}{\mathcal{T}^U}$ is an increasing (resp. decreasing) function of $X_{i,1}^*$ (resp. $X_{i,2}^*$) for each i, keeping all other entries in $\mathcal{D}_\infty^*$ fixed.

In detail, conditioned on the event of accepting H_1, the sample mean and empirical CDF for arm 1 are

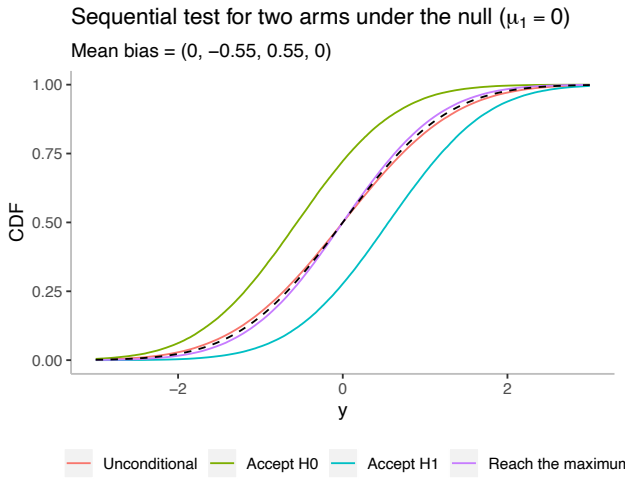

Figure 3.2: *Average of conditional empirical CDFs of arm 1 from repeated sequential tests for two arms as described in Section 3.2.2.*

positively and negatively biased, respectively. That is,

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\mu}_1(\mathcal{T}_M) \mid \mathcal{T}_M = \mathcal{T}^U\right] \\
&= \mathbb{E}\left[\widehat{\mu}_1(\mathcal{T}^U) \mid \mathcal{T}^U \leq \min\left\{\mathcal{T}^L, M\right\}\right] \geq \mu_1, \qquad (3.12) \\
&\sup_{y\in\mathbb{R}}\left(\mathbb{E}\left[\widehat{F}_{k,\mathcal{T}_M}(y) \mid \mathcal{T}_M = \mathcal{T}^U\right] - F_k(y)\right) \\
&= \sup_{y\in\mathbb{R}}\left(\mathbb{E}\left[\widehat{F}_{k,\mathcal{T}^U}(y) \mid \mathcal{T}^U \leq \min\left\{\mathcal{T}^L, M\right\}\right] - F_k(y)\right) \\
&\leq 0. \qquad (3.13)
\end{aligned}$$

Similarly, for arm 2, we have opposite signs of the sample mean and empirical CDF.

By the same reasoning, the signs of the conditional biases conditioned on the event of accepting H_0 are reversed:

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\mu}_1(\mathcal{T}_M) \mid \mathcal{T}_M = \mathcal{T}^L\right] \\
&= \mathbb{E}\left[\widehat{\mu}_1(\mathcal{T}^L) \mid \mathcal{T}^L \leq \min\left\{\mathcal{T}^U, M\right\}\right] \leq \mu_1, \qquad (3.14) \\
&\inf_{y\in\mathbb{R}}\left(\mathbb{E}\left[\widehat{F}_{k,\mathcal{T}_M}(y) \mid \mathcal{T}_M = \mathcal{T}^L\right] - F_k(y)\right) \\
&= \inf_{y\in\mathbb{R}}\left(\mathbb{E}\left[\widehat{F}_{k,\mathcal{T}^L}(y) \mid \mathcal{T}^L \leq \min\left\{\mathcal{T}^U, M\right\}\right] - F_k(y)\right) \\
&\geq 0. \qquad (3.15)
\end{aligned}$$

Thus, we conclude that, in all cases, the expected difference between the sample means are exaggerated toward "the direction of decision".

Remark 3.2. *These results hold regardless of whether the underlying distribution is under the null or alternative.*

Experiment: To demonstrate the conditional bias result, we set two standard normal arms with same means $\mu_1 = \mu_2 = 0$. In this experiment, we use upper and lower stopping boundaries based on naive

point-wise confidence intervals:

$$U(t) := z_{\alpha/2}\sqrt{\frac{2}{t}}, \quad \text{and} \quad L(t) = -U(t), \tag{3.16}$$

where α is set to 0.2 to show the bias better. Figure 3.2 show the conditional and unconditional biases of the empirical CDFs and sample means for arm 1 based on 10^5 repetitions of the experiment. The dashed line corresponds to the true underlying CDF. The red line refers to the average of the empirical CDFs, and the purple line corresponds to the average of the empirical CDFs conditioned on reaching the maximal time. Note that for these two cases, the empirical CDFs are neither positively nor negatively biased across all $y \in \mathbb{R}$.

However, for the cases corresponding to accepting H_1 (blue line) and accepting H_0 (green line), the empirical CDFs are negatively and positively biased, respectively: see inequalities (3.13) and (3.15). The conditional bias of the sample mean is also positive conditioning on the event of accepting H_1 and negative conditioning on the event of accepting H_0: see inequalities (3.12) and (3.14).

3.2.3 Best-arm identification algorithms

Suppose we have $K > 2$ sub-Gaussian arms with a common and known parameter σ^2. In many applications, we may want to identify which of the K arms has the largest mean parameter by using as few samples as possible.

In the previous subsection, we observed that, in the two-armed bandit setting, if we use a boundary-crossing compatible with a best-arm identification style algorithm with deterministic sampling, then the optimal arm has positive bias and the sub-optimal arm has a negative bias. We also showed that the same phenomenon happens for the best arm as chosen by the lil'UCB algorithm in Chapter 2. Recall that the lil'UCB algorithm consists of the following sampling, stopping and choosing rules [Jamieson et al., 2014]:

- Sampling: For $t = 1, \dots K$, draw a sample from each arm. For $t > K$, draw a sample from arm k if

$$k = \arg\max_{j \in [K]} \widehat{\mu}_j(t-1) + u^{\text{lil}}\left(N_j(t-1)\right),$$

  where $\beta, \epsilon, \lambda > 0$ and $\delta > 0$ are algorithm parameters and

$$\begin{aligned} u^{\text{lil}}(n) :=& (1+\beta)(1+\sqrt{\epsilon}) \\ & \times \sqrt{2\sigma^2(1+\epsilon)\log\left(\log((1+\epsilon)n)/\delta\right)/n}. \end{aligned}$$

- Stopping: the stopping time $\mathcal{T}$ is defined as the first time at which the following inequality holds for some $k \in [K]$:

$$N_k(t) \geq 1 + \lambda \sum_{j \neq k} N_j(t),$$

  where $\lambda > 0$ is an algorithm parameter.

- Choosing: $\kappa = \arg\max_{k \in [K]} N_k(\mathcal{T})$.

The following corollary to Theorem 3.1 shows a complementary result - the sample mean and the empirical CDF of a given arm are negatively and positively biased, respectively, conditioned on the event that the arm is not selected as the best arm. In contrast, on the same conditioning event, the empirical distribution of the selected arm is negatively biased.

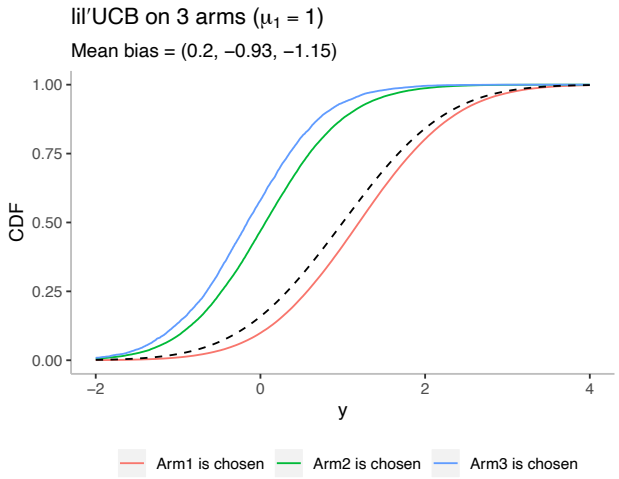

Figure 3.3: *Data from* 10^5 *lil'UCB algorithm runs on three unit-variance normal arms with* $\mu_1 = 1, \mu_2 = 0.5$ *and* $\mu_3 = 0$*, as described in Section 3.2.3.*

Corollary 3.3. *In the settings of the lil'UCB algorithm, for each* k *with* $\mathbb{P}(\kappa \neq k) > 0$*, we have that*

$$\mathbb{E}\left[\widehat{\mu}_k(\mathcal{T}) \mid \kappa \neq k\right] \leq \mu_k, \tag{3.17}$$

$$\inf_{y \in \mathbb{R}} \left(\mathbb{E}\left[\widehat{F}_{k,\mathcal{T}}(y) \mid \kappa \neq k\right] - F_k(y) \right) \geq 0. \tag{3.18}$$

Also, for each k *with* $\mathbb{P}(\kappa = k) > 0$*,*

$$\mathbb{E}\left[\widehat{\mu}_k(\mathcal{T}) \mid \kappa = k\right] \geq \mu_k, \tag{3.19}$$

$$\sup_{y \in \mathbb{R}} \left(\mathbb{E}\left[\widehat{F}_{k,\mathcal{T}}(y) \mid \kappa = k\right] - F_k(y) \right) \leq 0. \tag{3.20}$$

The proof of Corollary 3.3 can be found in Appendix B.1.2.

Remark 3.4. *Corollary 3.3 remains true for any other choice of* u^{lil} *if the function* $n \mapsto u^{lil}(n)$ *is decreasing.*

Experiment: To verify the previous claims, we conducted 10^5 trials of the lil'UCB algorithm on three unit-variance normal arms with $\mu_1 = 1, \mu_2 = 0.5$ and $\mu_3 = 0$. It is important to note that the signs of the biases do not depend on the choice of parameters or of the underlying distributions, but the magnitudes of the biases do. To best illustrates the bias results, we use an unusual set of algorithm parameters as $\delta = 0.2$, $\epsilon = 0.1$, $\beta = 0.5$ and $\lambda = 1$ in this experiment.

Figure 3.3 shows the averages of the empirical CDFs of arm 1 (the arm with the larges mean) conditioned on each arm being chosen as the best arm. The dashed line corresponds to the true underlying CDF. The red line, which lies below the true CDF, indicates that the empirical CDF of arm 1 conditioned on the event that the arm 1 is chosen as the best arm (i.e., $\kappa = 1$) is negatively biased; this then implies that the sample mean of the chosen arm is positively biased. In contrast, the green and blue lines, lying above the true CDF, show that conditioned on the event the arm 1 is not chosen as the best arm (i.e., $\kappa \neq 1$), the empirical CDF is positively biased and the sample mean is negatively biased.

Figure 3.4 displays the averages of the empirical CDFs of arm 2. Though arm 2 is not the best arm, we can check that the signs of the conditional biases follow the same pattern as arm 1. Conditioned on the

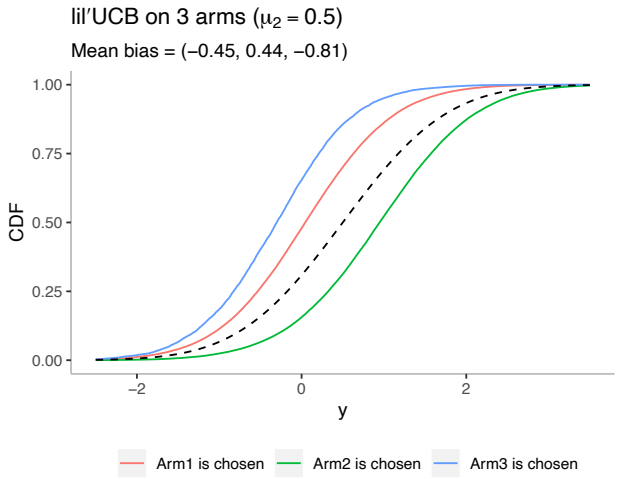

Figure 3.4: *Data collected by* 10^5 *lil'UCB algorithm runs on three unit-variance normal arms with* $\mu_1 = 1, \mu_2 = 0.5$ *and* $\mu_3 = 0$*, as described in Section 3.2.3.*

event that arm 2 is chosen as the best arm ($\kappa = 2$), the empirical CDF is negatively biased (green line), but conditioned on the event that arm 2 is not chosen as the best arm ($\kappa \neq 2$), the corresponding CDFs are now positively biased (red and blue lines), as expected.

3.3 Summary

In this chapter, we have investigated the sign of the conditional bias of monotone functions of the rewards in the MAB framework under an arbitrary conditioning event, generalizing the results in the previous chapter. In our analysis, we have exploited certain natural monotonicity properties of MAB experiments and have characterized the impact on the bias of both adaptivity in the data acquisition process and in the selection of the target for inference.

Several interesting extensions of our results are worth pursuing. We emphasize two important ones:

- It is still an open problem how to characterize the bias (conditional or unconditional) of other important functionals that are not necessarily monotone nor linear, such as the sample variance and sample quantiles.
- Several debiasing methods have been proposed in the MAB literature: see, e.g., Xu et al. [2013], Deshpande et al. [2018], Neel and Roth [2018], Nie et al. [2018], Hadad et al. [2019]. However, the existing approaches typically only adjust for adaptive sampling but ignore the other sources of adaptivity. Furthermore, they do not take account the natural direction of the bias, i.e., its sign. It is of interest to investigate how our results and techniques can be used in order to design more general debiased estimators.

Chapter 4

Consistency and bounds on bias and risks

4.1 Introduction

In previous two chapters, we conducted qualitative analyses of sample means in MABs. Specifically, we describe a simple monotonicity condition that determines the sign of the bias, including natural examples where it can be positively or negatively biased. Despite this progress, it is still obscure how large the bias is, and more generally, how the sample mean estimator behaves around the true mean.

In this chapter, we derive sufficient conditions under which the sample mean is consistent under all four aforementioned notions of adaptivity (sampling, stopping, choosing and rewinding, henceforth called the "fully adaptive setting"). Then, we study the magnitude of its bias and risk under general moment/tail conditions. Adaptive mean estimation, in each of the four senses described above, has received significant attention in both recent and older literature (only studied one at a time, not together). Below, we briefly discuss how our work relates to these past works, proceeding one notion at a time in approximate historical order.

We begin by noting that a single-armed bandit is simply a random walk, a setting where adaptive stopping has been extensively studied, since even the simplest of asymptotic questions are often nontrivial. For example, if a random walk is stopped at an increasing sequence of stopping times, the corresponding sequence of stopped sample means does *not* necessarily converge to μ, even in probability, without regularity conditions on the distribution and stopping rules (see Ch.1 of Gut [2009]). The book by Gut [2009] on stopped random walks is an excellent reference, beginning from the seminal paper of Wald and Wolfowitz [1948], and summarizing decades of advances in sequential analysis. Some relevant authors include Anscombe [1952], Richter [1965], Starr [1966], Starr and Woodroofe [1972], since they discuss inferential questions for stopped random walks or stopped tests, often in parametric and asymptotic settings. As far as we know, most of these results have not been extended to the MAB setting, which naturally involves adaptive sampling and choosing. Motivated by this, we provide new consistency results that hold in the fully adaptive setting.

Next, for the problem of estimation following a sequential test, Cox [1952] and Siegmund [1978] developed an asymptotic expression for the size of the bias of the sample mean. Further, the moment bounds derived in de la Pena et al. [2004], Peña et al. [2008] for self-normalized processes can be converted into bounds, on the ℓ_p-risk of the sample mean. However, both sets of results apply only for a fixed arm (since they work in the one-armed setting), for a specific stopping rule (a sample mean crossing a boundary) and for a restricted class of arms (in our context, sub-Gaussian arms) and do not directly apply to adaptively chosen arms in an MAB setting, unlike the results that we derives.

Third, the recent literature on best-arm identification in MABs has often used anytime uniform concen-

tration bounds for the sample mean of each arm around its true mean [Jamieson et al., 2014, Kaufmann et al., 2016] also known as finite-LIL bounds. Historically, these were called confidence sequences and were developed by Darling and Robbins [1967a, 1968], Lai [1976], though both theoretical and practical advances outside the MAB literature have been made recently [Balsubramani, 2014, Balsubramani and Ramdas, 2016, Howard et al., 2018b]. While these results yield high probability deviation inequalities that allow for adaptive rewinding, they cannot be immediately converted into bias and risk bounds. Below, we develop variants of these bounds and incorporate them into our risk analysis to to cover all the cases in the fully adaptive setting.

Last, Russo and Zou [2016] recently derived information-theoretic bounds for the selection bias introduced by adaptive choosing. This work, soon extended by Jiao et al. [2017], showed that if a fixed number of samples is collected from each distribution, then the bias (or expected ℓ_2 loss) of the sample mean of an adaptively chosen arm can be bounded using the mutual information between the arm index and the observed data. From our MAB perspective, these bounds only hold for a deterministic sampling rule that is stopped at a fixed time. We derives new bias and risk bounds based on the mutual information for the fully adaptive setting.

In sum, characterizing the risk and bias under all four notions of adaptivity simultaneously is an interesting and challenging problem. Below, we summarize our contributions and describe the organization of this chapter:

1. We formulate sufficient conditions for consistency of a sequence of sample means in the fully adaptive setting which only require the existence of a finite mean for each arm (Proposition 4.2).
2. For the ℓ_2 loss and for arms with finite moments, we derive risk bounds for the sample mean in the fully adaptive setting that includes an adaptive arm choice and adaptive rewinding (Theorem 4.6 and Corollary 4.8).
3. By considering certain Bregman divergences between the sample and true mean as loss functions and for arms with exponentially decaying tails, we derive sharp risk bounds for a fixed target at a stopping time (Theorem 4.12) which are in turn used to derive quantitative upper and lower bounds for the bias under adaptive sampling and stopping (Corollary 4.14).
4. Under the fully adaptive setting including adaptive arm choice and adaptive rewinding, we show that by inducing a small "adaptive normalizing factor" in a log-log scale, we can extend the above results to derive bounds on the normalized risk of the sample mean to the fully adaptive setting (Theorem 4.16 and Corollary 4.19).

4.2 Consistency of the sample mean

In sequential data analysis we often estimate the mean not just once but many times as new data become available. Let $\tau_1 \leq \tau_2 \leq \cdots$ be a sequence of non-decreasing random times, and thus $N_k(\tau_1) \leq N_k(\tau_2) \leq \cdots$. A natural question is to identify conditions under which the sample mean $\widehat{\mu}_{\kappa_t}(\tau_t)$ is consistent, in the sense that the sequence $\widehat{\mu}_{\kappa_t}(\tau_t) - \mu_{\kappa_t}$ converges to zero, almost surely or in probability, as $t \to \infty$.

It is well known that the condition $\mathbb{E}\left[N_k(\tau_t)\right] \to \infty$ as $t \to \infty$ is not sufficient to guarantee consistency of the sample mean even for a fixed target arm k, as demonstrated in the next example.

Example 4.1. *Let P_1 and P_2 be standard normal distributions. Set $\nu_1(1) = 1$, that is, the algorithm always picks the first distribution at $t = 1$. For $t \geq 2$, set $A_t = \mathbb{1}(|Y_1| > z_{\alpha/2}) + 1$ where z_α is the α-upper quantile of the standard normal which means that we pick a single (random) arm forever based on the first observation Y_1. Finally, let $t^* \geq 2$ be a deterministic stopping time. Then, we have*

$$\mathbb{E}N_1(t^*) = 1 + (t^* - 1)(1 - \alpha) \longrightarrow \infty \text{ as } t^* \longrightarrow \infty.$$

Note however that

$$\mathbb{P}\left(\widehat{\mu}_1(t^*) > z_{\alpha/2}\right) \geq \mathbb{P}\left(|Y_1| > z_{\alpha/2}\right) = \alpha, \ \ \forall t^* \geq 2.$$

Therefore $\mathbb{P}\left(\widehat{\mu}_1(t^*) > z_{\alpha/2}\right)$ *does not converge to zero even if* t^* *and* $\mathbb{E}N_1(t^*)$ *approach infinity, and hence* $\widehat{\mu}_1(t^*)$ *does not converge to the true mean* $\mu_1 = 0$ *in probability.*

For each fixed $k \in [K]$, Theorem 2.1 in Gut [2009] immediately yields that

$$\text{if } \ N_k(\tau_t) \xrightarrow{a.s.} \infty \ \text{ as } \ t \to \infty, \ \ \text{then } \ \widehat{\mu}_k(\tau_t) \xrightarrow{a.s.} \mu_k \ \text{ as } \ t \to \infty. \tag{4.1}$$

Theorem 2.2 in Gut [2009] further implies that, in the previous display (4.1), we can replace almost sure convergence with the convergence in probability in both the condition and conclusion. In our next result, we generalize these claims to the multi-armed setting with an adaptively chosen arm. Note that here we only need the underlying distributions to have finite first moments.

Proposition 4.2. *The following statements hold for any sequence of choice functions* $\kappa_t : \mathcal{D}(\tau_t) \to [K]$ *that are based on data up to time* τ_t*:*

$$\textit{if } \ N_{\kappa_t}(\tau_t) \xrightarrow{a.s.} \infty \ \textit{ as } \ t \to \infty, \ \ \textit{then } \ \widehat{\mu}_{\kappa_t}(\tau_t) - \mu_{\kappa_t} \xrightarrow{a.s.} 0 \ \textit{ as } \ t \to \infty, \ \textit{and} \tag{4.2}$$

$$\textit{if } \ N_{\kappa_t}(\tau_t) \xrightarrow{p} \infty \ \textit{ as } \ t \to \infty, \ \ \textit{then } \ \widehat{\mu}_{\kappa_t}(\tau_t) - \mu_{\kappa_t} \xrightarrow{p} 0 \ \textit{ as } \ t \to \infty. \tag{4.3}$$

The proof of the proposition is deferred to Appendix C.5.1.

Since $|\widehat{\mu}_{\kappa_t}(\tau_t) - \mu_{\kappa_t}| = \sum_{k=1}^{K} \mathbb{1}\left(\kappa_t = k\right) |\widehat{\mu}_k(\tau_t) - \mu_k|$, if we instead assume the stronger condition that $N_k(\tau_t) \to \infty$ for all $k \in [K]$, almost surely or in probability, Theorem 2.1 and 2.2 in Gut [2009] immediately imply consistency of the sample means. However, the following example demonstrates that even if the number of draws for each *fixed* arm does not converge to infinity, Proposition 4.2 can guarantee the consistency of the chosen sample mean which, in contrast, cannot be directly implied by Theorem 2.1 and 2.2 in Gut [2009].

Example 4.3. *Let* P_1 *and* P_2 *be two identical continuous distributions with finite means. Set* $\nu_1(1) = 1$ *and* $\nu_2(2) = 1$*, meaning that we begin by sampling each arm once. For all times* $t \geq 3$*, we set* $A_t = \mathbb{1}(Y_1 > Y_2) + 1$*, meaning that we pick a single (random) arm forever. Finally, let* $t^* \geq 3$ *be a deterministic stopping time. The number of draws from each arm does not diverge to infinity either almost surely or in probability as* $t^* \to \infty$ *since for* $k = 1, 2$*, we have*

$$\mathbb{P}\left(N_k(t^*) \leq 1\right) = \frac{1}{2}, \ \ \forall t^* \geq 3.$$

Now, let $\kappa = \mathbb{1}\left(N_1(t^*) \leq N_2(t^*)\right) + 1$*, that is, we choose the arm with more data when we stop. Then, the number of draws from the chosen arm is always equal to* $t^* - 1$ *and the sufficient condition in Proposition 4.2 is trivially satisfied. Thus, even though* $N_k(t^*)$ *does not diverge to* ∞ *(almost surely or in probability) for any fixed* k*, our proposition still guarantees that* $\widehat{\mu}_\kappa(t^*) - \mu_\kappa \xrightarrow{a.s.} 0$ *as* $t^* \to \infty$.

The above example demonstrates the additional subtlety in the conditions for consistency when moving from a fixed arm to an adaptively chosen one.

4.3 Risk of sample mean under arms with finite moments

Having established that the sample means are consistent estimators of the true means of the arms, in the subsequent sections, we will turn to the more challenging tasks of deriving finite sample bounds on the magnitude of both the bias and risk under different nonparametric assumptions on the arms. Specifically,

we will be concerned with two notions of ℓ_2 risk for the sample mean estimator: the classic or *unnormalized* one, corresponding to the squared error loss and given by

$$\text{[Unnormalized } \ell_2 \text{ risk]} \quad \mathbb{E}\left[\left(\widehat{\mu}_\kappa(\tau) - \mu_\kappa\right)^2\right], \tag{4.4}$$

and a weighted or *normalized* variant, defined as

$$\text{[Normalized } \ell_2 \text{ risk]} \quad \mathbb{E}\left[N_\kappa(\tau)\left(\widehat{\mu}_\kappa(\tau) - \mu_\kappa\right)^2\right]. \tag{4.5}$$

As we will see shortly, the unnormalized risk is a function of both sampling and stopping rules, while the normalized risk is upper bounded by a term that only depends on the choosing rule. The two types of risk bounds are rather different in both their form and interpretability, and each elucidate complementary aspects of the problems. In addition to normalized and unnormalized ℓ_2 bounds, we will also give analogous ℓ_1 bounds.

For each $p \geq 1$ and $k \in [K]$, we define the centered p-norm of arm k as

$$\sigma_k^{(p)} := \left(\int |x - \mu_k|^p \, \mathrm{d}P_k(x)\right)^{1/p}. \tag{4.6}$$

From Jensen's inequality, we know that if $p_1 \leq p_2$ then $\sigma_k^{(p_1)} \leq \sigma_k^{(p_2)}$ for each $k \in [K]$. With a slight abuse of notation, below we will denote the standard deviation of the k-th arm with σ_k instead of $\sigma_k^{(2)}$.

4.3.1 Unnormalized and normalized ℓ_2 risks under nonadaptive sampling and stopping

Recall that if the sampling (or stopping) rule is independent of the realizations of the arms, we call it a nonadaptive sampling (or stopping) rule, respectively. Under nonadaptive sampling and stopping rules, the unnormalized ℓ_2-risk of the sample mean for arm k is given by

$$\begin{aligned} \mathbb{E}\left(\widehat{\mu}_k(T) - \mu_k\right)^2 &= \mathbb{E}\left[\mathbb{E}\left[\left(\widehat{\mu}_k(T) - \mu\right)^2 \mid \{A_t\}_{t=1}^T\right]\right] \\ &= \mathbb{E}\left[\frac{\sigma_k^2}{N_k(T)}\right], \end{aligned} \tag{4.7}$$

where the second equality comes from the independence assumption on the sampling nd stopping rules and the fact that $\mathbb{E}\left(\widehat{\mu}_k(n) - \mu_k\right)^2 = \frac{\sigma_k^2}{n}$ where $\widehat{\mu}_k(n)$ is a sample average of n i.i.d. observations from a distribution with mean μ_k and variance σ_k^2. Next, define the effective sample size for arm k as

$$n_k^{\text{eff}} := \left[\mathbb{E}\left(1/N_k(T)\right)\right]^{-1}. \tag{4.8}$$

Then, under nonadaptive sampling and stopping, the ℓ_2 risk of the sample mean for arm k can immediately be derived to equal

$$\mathbb{E}\left(\widehat{\mu}_k(T) - \mu_k\right)^2 = \frac{\sigma_k^2}{n_k^{\text{eff}}}.$$

Clearly, the effective sample size n_k^{eff} depends on both the nonadaptive sampling and stopping rules, as it quantifies the combined effects of these rules on the ℓ_2 risk of $\widehat{\mu}_k(T)$. In contrast, the normalized risk is agnostic to the choices of such rules. In detail, we show next that the minimax normalized ℓ_2 risk for estimating the mean of the kth arm over all nonadaptive data collection procedures is σ_k^2, and this risk value is achieved by the sample mean.

Proposition 4.4. *For any fixed $k \in [K]$, let $\mathbb{P}_k(\mu_k, \sigma_k)$ be the class of distributions on an arm k with mean μ_k and variance σ_k^2. Let $\mathbb{V}$ and $\mathbb{T}$ be classes of nonadaptive sampling and stopping rules satisfying $N_k(T) \geq 1$. Finally, let $Q = Q(P_k, \nu, T)$ be the induced distribution on observations from the arm k with P_k distribution under nonadaptive sampling ν and stopping T. Then, the minimax normalized ℓ_2 risk is given by*

$$\inf_{\widehat{\mu}_k} \sup_{\substack{P_k \in \mathbb{P}_k(\mu_k, \sigma_k) \\ \nu \in \mathbb{V}, T \in \mathbb{T}}} \mathbb{E}_Q \left[N_k(T) \left(\widehat{\mu}_k(T) - \mu_k \right)^2 \right] = \sigma_k^2, \tag{4.9}$$

where the infimum is over all estimators. Furthermore, for any given $P_k \in \mathbb{P}_k(\mu, \sigma_k^2)$, $\nu \in \mathbb{V}$ and $T \in \mathbb{T}$, the sample mean estimator achieves the minimax risk.

The proof of the proposition is based on standard decision-theoretic arguments and can be found in Appendix C.5.2.

In practice, we often do not know ahead of time which arm k would be the most interesting to study before looking at the data. For instance, we may want to estimate the mean for the arm with the largest observed empirical mean, or the second largest, or even the smallest. In this case, the target of inference is μ_κ, where κ is an adaptive choice which possibly depend on the collected data $\mathcal{D}_\mathcal{T}$.

Following Jiao et al. [2017], to quantify dependence between κ and $\mathcal{D}_\mathcal{T}$, we adopt an information theoretic dependence measure (f_q-divergence):

$$I_q(\kappa; \mathcal{D}_\mathcal{T}) := D_{f_q} \left(P_{(\kappa, \mathcal{D}_\mathcal{T})} | P_\kappa \otimes P_{\mathcal{D}_\mathcal{T}} \right), \tag{4.10}$$

where $q \geq 1$, $f_q(x) := |x - 1|^q$ and $D_{f_q}(Q'|Q) := \int f_q \left(\frac{\mathrm{d}Q'}{\mathrm{d}Q} \right) \mathrm{d}Q$, assuming that $Q' \ll Q$. It can be easily checked that $I_q(\kappa; \mathcal{D}_\mathcal{T}) \geq 0$ and that $I_q(\kappa; \mathcal{D}_\mathcal{T}) = 0$ if and only if κ and $\mathcal{D}_\mathcal{T}$ are independent. It can be also showed that $I_q(\kappa, \mathcal{D}_\mathcal{T})$ can be upper bounded as

$$I_q(\kappa, \mathcal{D}_\mathcal{T}) \leq 1 + \sum_{k=1}^{K} p_k^2 \left(\left| \frac{1}{p_k} - 1 \right|^q - 1 \right),$$

where $p_k := \mathbb{P}(\kappa = k), \ \ \forall k \in [K]$. In particular,

$$I_q(\kappa, \mathcal{D}_\mathcal{T}) \leq \frac{K-1}{K} \left[(K-1)^{q-1} + 1 \right] < 1 + (K)^{q-1},$$

for $1 \leq q \leq 2$ [see Jiao et al., 2017, Lemma 1].

For nonadaptive sampling and stopping (and hence $\mathcal{T} = T$), Jiao et al. [2017] showed how to bound the bias of adaptively chosen random variables with finite moments by using $I_q(\kappa; \mathcal{D}_T)$. More precisely, suppose each $\widehat{\mu}_k - \mu_k$ has zero mean and its p-norm is given by $\left(\mathbb{E} \left(\widehat{\mu}_k - \mu_k \right)^p \right)^{1/p} = \sigma_k^{(p)}$. Also, for any $r \geq 1$, let $\| \sigma_\kappa^{(p)} \|_r$ be the r-norm of $\sigma_\kappa^{(p)}$, naturally defined by

$$\| \sigma_\kappa^{(p)} \|_r = \left(\sum_{k=1}^{K} \mathbb{P}(\kappa = k) \left(\sigma_k^{(p)} \right)^r \right)^{1/r}. \tag{4.11}$$

Then, Jiao et al. [2017] proved that, for any $p, q > 1$ with $1/p + 1/q = 1$, the bias of $\widehat{\mu}_\kappa$ can be bounded as

$$|\mathbb{E}\widehat{\mu}_\kappa - \mu_\kappa| \leq \| \sigma_\kappa^{(p)} \|_p I_q^{1/q}(\kappa, \mathcal{D}_T). \tag{4.12}$$

This result can be extended to a bound on the ℓ_2 risk of an adaptively chosen sample mean (under nonadaptive sampling and stopping) as follows.

Proposition 4.5. *Consider some nonadaptive sampling and stopping rules, and assume that each arm has a finite $2p$-norm for a given $p > 1$. Then, the normalized ℓ_2 risk of the sample mean can be bounded as*

$$\mathbb{E}\left[N_\kappa(T)\left(\widehat{\mu}_\kappa(T) - \mu_\kappa\right)^2\right] \leq \|\sigma_\kappa\|_2^2 + C_p \left\|\sigma_\kappa^{(2p)}\right\|_{2p}^2 I_q^{1/q}(\kappa, \mathcal{D}_T), \tag{4.13}$$

where C_p is a constant depending only on p and $q > 1$ is such that $1/p + 1/q = 1$.

The proof of Proposition 4.5 can be found in Appendix C.5.3 and is based on a variational representation of the f_q-divergence along with the Marcinkiewicz–Zygmund inequality [Marcinkiewicz and Zygmund, 1937]. Note that if $\mathbb{P}(\kappa = k) = 1$ then the mutual dependence term $I_q^{1/q}(\kappa, \mathcal{D}_\mathcal{T})$ is equal to 0 and we recover the exact ℓ_2 risk σ_k^2. Similarly, if the selected arm K is chosen in a random but nonadaptive manner, we have that $I_q^{1/q}(K, \mathcal{D}_\mathcal{T}) = 0$ and thus, the bound in (4.13) reduces to $\|\sigma_K\|_2^2$.

4.3.2 Normalized ℓ_2 risk and unnormalized ℓ_1 risk under fully adaptive settings

The techniques used in the previous section deliver risk bounds only under nonadaptive sampling and stopping rules but they do not generalize readily to fully adaptive settings. In particular, the bias bound in Jiao et al. [2017] and the risk bound in Proposition 4.5 are not directly applicable because each $\widehat{\mu}_k(\tau) - \mu_k$ is no longer centered, due to the bias caused by adaptive sampling, stopping and rewinding. Furthermore, the bound for the bias given in equation (4.12) no longer holds under the fully adaptive setting because the bias can be non-zero even if κ is independent of $\mathcal{D}$.

Below we show that the normalized ℓ_2 risk bound for nonadaptive sampling and stopping strategies given in Proposition 4.5 generalizes to the fully adaptive setting, assuming the existence of higher moments and with a slightly stronger risk normalization factor of $N_k(\tau)/\log N_k(\tau)$, which can be regarded as a (small) price for adaptivity.

For any t such that $N_k(t) > 1$ for all $k \in [k^*]$, we define

$$\widetilde{N}_k(t) := \frac{N_k(t)}{\log N_k(t)}, \quad k \in [K]. \tag{4.14}$$

We now present the main result of this section.

Theorem 4.6. *Suppose each arm has a finite $2(p+\epsilon)$-norm for some $p \geq 1$ and $\epsilon > 0$. Consider any adaptive sampling rule and stopping time $\mathcal{T}$ such that $\min_{k\in[K]} N_k(\tau) \geq 3$ almost surely for an adaptive rewound time $\tau \leq \mathcal{T}$. Then, for any adpatively chosen arm κ, it holds that*

$$\mathbb{E}\left[\widetilde{N}_\kappa(\tau)\left(\widehat{\mu}_\kappa(\tau) - \mu_\kappa\right)^2\right] \leq C_{1,\epsilon}\|\sigma_\kappa\|_2^2 + C_{p,\epsilon}\|\sigma_\kappa\|_{2p}^2 I_q^{1/q}\left(\kappa, \mathcal{D}_\mathcal{T}\right), \tag{4.15}$$

where $q > 1$ satisfies $1/p + 1/q = 1$ and $C_{1,\epsilon}$ is a positive constant depending only on ϵ, and $C_{p,\epsilon}$ is a positive constant depending only on p, ϵ.

Compared to the nonadaptive risk bound in Proposition 4.5, the bound (4.15) under the fully adaptive setting only suffers a multiplicative logarithmic normalization term $\log N_\kappa(\tau)$ under slightly stronger moment condition. It is also important to note that the bound (4.15) depends on the second moment terms $\{\sigma_k\}_{k=1}^K$ only, and not on any higher moment.

The proof of Theorem 4.6, given in Section C.2.2, combines the novel deviation inequality for the normalized ℓ_2 loss of Lemma 4.7 below, which holds for a fixed arm k, with the variational representation of the f_q-divergence which handles adaptive choosing.

Lemma 4.7. *Consider some adaptive sampling and stopping rules. For a fixed arm $k \in [K]$ with a finite $2p$-norm, where $p > 1$, and any random time τ such that $N_k(\tau) \geq 3$ almost surely, it holds that, for any $\delta \geq 0$,*

$$\mathbb{P}\left(\widetilde{N}_k(\tau)\left(\frac{\widehat{\mu}_k(\tau)-\mu_k}{\sigma_k}\right)^2 \geq \delta\right) \leq \frac{C_p}{\delta^p}, \tag{4.16}$$

where C_p is a constant depending only on p.

We believe this is the first polynomially decaying tail bound on the ℓ_2 risk of the sample mean that holds at any arbitrary random time and only assuming arms with finite first $2p$ moments. This inequality is thus possibly of independent interest; its proof is based on the ℓ_p-version of the Dubins-Savage inequality given by Khan [2009]; see Appendix C.2.1.

Now, for any $r > 0$, define the r-th order logarithmically discounted sample size of an adaptively chosen arm as

$$\widetilde{n}_\kappa^{\mathrm{eff},r} := \left[\mathbb{E}\left[1/\widetilde{N}_\kappa^r(\tau)\right]\right]^{-1/r}, \quad r > 0, \tag{4.17}$$

where the expectation is over the randomness in all four sources of adaptivity. This quantity is nonrandom, and the subscript κ merely differentiates it from the effective sample size of a fixed arm, and is not to be interpreted as residual randomness. It is easy to check that $\widetilde{n}_\kappa^{\mathrm{eff},r}$ is decreasing with respect to r, by Jensen's inequality. The following corollary provides bounds for the unnormalized ℓ_{2r} risk of the sample mean for all $r \in (0,1)$ based on $\widetilde{n}_\kappa^{\mathrm{eff},r}$. The proof of the corollary can be found in Appendix C.2.3.

Corollary 4.8. *Suppose each arm has a finite $2(p+\epsilon)$-norm for some $p \geq 1$ and $\epsilon > 0$. Then, for any $r \in (0,1)$, the r-quasi-norm of the ℓ_2-loss is bounded as*

$$\left[\mathbb{E}\left(\widehat{\mu}_\kappa(\tau)-\mu_\kappa\right)^{2r}\right]^{1/r} \leq \frac{C_{1,\epsilon}\|\sigma_\kappa\|_2^2 + C_{p,\epsilon}\|\sigma_\kappa\|_{2p}^2 I_q^{1/q}(\kappa, \mathcal{D}_\mathcal{T})}{\widetilde{n}_\kappa^{\mathrm{eff},r/(1-r)}}. \tag{4.18}$$

In particular, by choosing $r = 1/2$, the above results immediately yields a bound for the ℓ_1 risk:

$$\mathbb{E}\left|\widehat{\mu}_\kappa(\tau)-\mu_\kappa\right| \leq \sqrt{\frac{C_{1,\epsilon}\|\sigma_\kappa\|_2^2 + C_{p,\epsilon}\|\sigma_\kappa\|_{2p}^2 I_q^{1/q}(\kappa, \mathcal{D}_\mathcal{T})}{\widetilde{n}_\kappa^{\mathrm{eff},1}}}. \tag{4.19}$$

Note that if the choosing rule κ is equal to k (so that $I_q(\kappa, \mathcal{D}_\mathcal{T}) = 0$), the ℓ_1 risk bound (4.19) matches to the nonadaptive standard ℓ_1 risk bound, of order of $\sigma_k/\sqrt{n}$, with the sample size n replaced by the logarithmically discounted effective sample size $\widetilde{n}_k^{\mathrm{eff},1}$. In Section 4.4.2, we derive an alternative bound that depend on the undiscounted effective sample size n_k^{eff}, for a fixed target arm and at a stopping time, by assuming stronger tail conditions.

One may wonder whether the logarithmic discounting factor in the normalized risk is necessary to derive a finite upper bound. For arms with finite variance, we can show that in general there is no finite upper bound on the normalized risk $\mathbb{E}\left[N_k(\mathcal{T})\left(\widehat{\mu}_k(\mathcal{T})-\mu_k\right)^2\right]$ by using the following example.

Example 4.9. *Suppose each arm has a finite variance σ_k^2. For a fixed k, assume that $N_k(t) \to \infty$ almost surely as $t \to \infty$. For any $b \geq 3$, we define the following stopping rule.*

$$\mathcal{T}_b := \inf\left\{t \geq 1 : N_k(t) \geq b \text{ and } \frac{S_k(t) - \mu_k N_k(t)}{\sigma_k\sqrt{N_k(t)\log\log N_k(t)}} \geq 1\right\}. \tag{4.20}$$

Due to the law of the iterated logarithm, we know that $\mathbb{P}(\mathcal{T}_b < \infty) = 1$. From the definition of $\mathcal{T}_b$, we immediately infer that

$$\sigma_k^2\mathbb{E}\left[\log\log N_k(\mathcal{T}_b)\right] \leq \mathbb{E}\left[N_k(\mathcal{T}_b)\left(\widehat{\mu}_k(\mathcal{T}_b)-\mu_k\right)^2\right]. \tag{4.21}$$

Since the left hand side approaches infinity as $b \to \infty$, we see that there is no finite upper bound on the normalized risk $\mathbb{E}\left[N_k(\mathcal{T})\left(\widehat{\mu}_k(\mathcal{T}) - \mu_k\right)^2\right]$ in general.

The above example demonstrates that some correction to the normalized risk, such as the logarithmic discounting of the sample size in (4.14), is necessary to derive a finite risk bound like in Theorem 4.6. It is unclear whether the logarithmic discounting we used is optimal or if a smaller factor would have sufficed.

In the next section we show that, for arms with exponentially decaying tails, we can deploy a smaller discounting factor, measured on a log-log scale, that leads to upper and lower bounds matching up to a constant term; see Theorem 4.16.

4.4 Risk bounds for arms with exponential tails

In this section we will assume stronger tail-decaying conditions on the arms and derive risk bounds for the sample means under various degree of adaptivity. While the analysis and results might be cleanest for the mean-squared error of sub-Gaussian distributions (as is commonly assumed in the bandit literature), the proof for the more general case involving Bregman risks of sub-ψ distributions follows the same line of argument and hence we choose to present the results in a unified way. The sub-Gaussian results can be easily inferred as a special case.

4.4.1 Sub-ψ arms and Bregman divergences as loss functions

For fixed numbers $\lambda_{\min} < 0 < \lambda_{\max}$, let $\Lambda = (\lambda_{\min}, \lambda_{\max}) \subseteq \mathbb{R}$ be an open interval that contains 0. A function $\psi : \Lambda \to [0, \infty)$ is called CGF-like if it obeys natural properties of a cumulant generating function (CGF), specifically that it is a non-negative, twice-continuously differentiable and strictly convex function $\psi(0) = \psi'(0) = 0$.

A probability distribution P is called *sub-ψ* if the CGF of the centered distribution exists and is equal to or upper bounded by a "CGF-like" function ψ, that is,

$$\log \mathbb{E}_{Y\sim P}[e^{\lambda(Y-\mu)}] \le \psi(\lambda), \quad \forall \lambda \in \Lambda \subseteq \mathbb{R}. \tag{4.22}$$

This assumption is quite general and applies to all distributions with a CGF, including natural exponential family distributions, sub-Gaussian and sub-exponential distributions. Throughout this section, we assume each arm is in a sub-ψ class unless otherwise specified.

Our analyses make frequent use of $\psi_\mu^* : \Lambda^* \to \mathbb{R}$, the convex conjugate of $\psi_\mu(\lambda) := \lambda\mu + \psi(\lambda)$ defined as

$$\psi_\mu^*(z) := \sup_{\lambda\in\Lambda} \lambda z - \psi_\mu(\lambda), \quad \forall z \in \Lambda^* := \left\{ x \in \mathbb{R} : \sup_{\lambda\in\Lambda} \lambda x - \psi_\mu(\lambda) < \infty \right\}. \tag{4.23}$$

For arms in a sub-ψ class, it turns out to be natural to define the loss function as the Bregman divergence with respect to ψ_μ^*:

$$D_{\psi_\mu^*}(\widehat{\mu}, \mu) = \psi_\mu^*(\widehat{\mu}) - \psi_\mu^*(\mu) - \psi_\mu^{*\prime}(\mu)\,(\widehat{\mu} - \mu)\,. \tag{4.24}$$

For instance, if the underlying distribution is sub-Gaussian, then the Bregman divergence reduces to the scaled ℓ_2 loss. For more examples, see Appendix C.1. More generally, the Bregman divergence is equivalent to the KL loss when the underlying distribution is a natural univariate exponential family with a density

$$p_\theta(x) = \exp\{\theta x - B(\theta)\}, \quad \theta \in \Theta \subset \mathbb{R},$$

with respect to a reference measure γ, where $\Theta \subset \left\{\theta \in \mathbb{R} : \int e^{\theta x}\gamma(\mathrm{d}x) < \infty\right\}$ is the natural parameter space and $B\colon \Theta \to \mathbb{R}$ is a strictly convex function given by $\theta \mapsto \int e^{\theta x}\gamma(\mathrm{d}x)$. We assume throughout that Θ is nonempty and open.

For a fixed $\theta \in \Theta$, define $\Lambda_\theta := \{\lambda \in \mathbb{R} : \lambda + \theta \in \Theta\}$ and, for each $\lambda \in \Lambda_\theta$, let $\psi(\lambda) = \psi(\lambda;\theta) := B(\lambda+\theta) - B(\theta) - \lambda B'(\theta)$. Using the properties of the log-partition function B, it can be easily checked that p_θ is sub-ψ. Since B is strictly convex, there is a one-to-one correspondence between the natural parameter space and the mean value parameter space $\mathrm{M} = \{\mu \in \mathbb{R}\colon \mu = B'(\theta), \theta \in \Theta\}$. For any μ_0, μ_1 in the mean parameter space, let θ_0, θ_1 be corresponding natural parameters. The KL divergence between p_{θ_1} and p_{θ_0} induces a natural loss between μ_1 and μ_0 which is often called the KL loss:

$$\ell_{KL}(\mu_1, \mu_0) := D_{KL}\left(p_{\theta_1} \| p_{\theta_0}\right).$$

The following well-known fact, based on the properties of the CGF of an exponential family and the duality of Bregman divergence, formally captures how the KL loss is related to the Bregman loss. For completeness, we present a proof in Appendix C.5.4.

Fact 4.10. *Let ψ be the CGF of a centered distribution in a one-dimensional exponential family. Then, for any μ_1 and μ_0 in the mean parameter space, we have*

$$\ell_{KL}(\mu_1, \mu_0) = D_{\psi^*_{\mu_0}}(\mu_1, \mu_0) = \psi^*_{\mu_0}(\mu_1) = \psi^*(\mu_1 - \mu_0). \tag{4.25}$$

Further, the last two equalities hold for any CGF-like ψ.

Since the identity (4.25) recovers the ℓ_2 loss for sub-Gaussian arms and the KL loss for exponential family arms, the Bregman divergence $D_{\psi^*_{\mu_\kappa}}(\widehat{\mu}_\kappa(\tau), \mu_\kappa)$ is a natural loss function for the mean value parameter when the arms are sub-ψ.

Below, we will show that in the deterministic setting where a fixed number n of independent observations are drawn from a single fixed distribution, the minimax Bregman risk for distributions belonging to an exponential family is of order $\frac{1}{n}$, and that the sample mean is minimax rate-optimal. To get a lower bound, we need an additional regularity condition on the loss function. For any function $d : \mathrm{M} \times \mathrm{M} \to [0, \infty)$, we say that d satisfies the *local triangle inequality condition* [Yang and Barron, 1999] if there exist positive constants $M \le 1$ and ϵ_0 such that for any $\mu_0, \mu_1, \mu_2 \in \mathrm{M}$, if $\max\{d(\mu_1, \mu_0), d(\mu_2, \mu_0)\} \le \epsilon_0$, then $d(\mu_1, \mu_0) + d(\mu_2, \mu_0) \ge M \max\{d(\mu_1, \mu_2), d(\mu_2, \mu_1)\}$. The local triangle inequality condition is satisfied by the square root KL divergence between Gaussian distributions with $M = 1$. For general exponential family distributions, we may restrict the parameter space to make the condition satisfied. In particular, if $\inf_{\theta\in\Theta} B''(\theta) > 0$ and $\sup_{\theta\in\Theta} B''(\theta) < \infty$, the condition satisfied with $M = \sqrt{\frac{\inf_{\theta\in\Theta} B''(\theta)}{\sup_{\theta\in\Theta} B''(\theta)}} \in (0, 1)$.

Under the local triangle inequality condition, we can prove that the minimax rate of convergence is $\frac{1}{n}$ and it can be achieved by the sample mean. The proof can be found in Appendix C.5.5. Note that, for the sub-Gaussian case, the risk reduces to the normalized ℓ_2 risk we studied in the previous section.

Proposition 4.11. *Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample from a distribution in a natural exponential family $\{p_\theta : \theta \in \Theta\}$. For each $\theta \in \Theta$, let μ be the mean parameter and ψ_μ is the cumulant generating function corresponding to θ. Then the risk of the sample mean, $\widehat{\mu}(n) = \frac{1}{n}\sum_{i=1}^n X_i$, is bounded as*

$$\mathbb{E}_{P_\theta}\left[n D_{\psi^*_\mu}(\widehat{\mu}(n), \mu)\right] \le 2, \quad \forall \theta \in \Theta. \tag{4.26}$$

Also, if $\sqrt{D_{\psi^*_\mu}}$ satisfies the local triangle inequality condition, then, for a large enough n, the minimax risk is lower bounded as

$$\inf_{\widehat{\mu}} \sup_{\theta\in\Theta} \mathbb{E}_{P_\theta}\left[n D_{\psi^*_\mu}(\widehat{\mu}(n), \mu)\right] \ge \frac{M \log 2}{16}. \tag{4.27}$$

Proposition 4.11 provides the inspiration for the results of the subsequent sections, where we will establish various upper bounds on both normalized and unnormalized versions the Bregman divergence risk under various degrees of adaptivity. Specifically, starting with the simple settings of a fixed target arm at a stopping time, we derive a tight upper bound on the unnormalized Bregman risk based on the effective sample. Then we move to the fully adaptive setting and show that by inducing a small "adaptive normalizing factor" in a log-log scale, we can extend the bound (4.26) on the normalized risk of the sample mean to the fully adaptive setting.

4.4.2 Bregman divergence risk bounds for a fixed target arm at a stopping time

Recall that for each $k \in [K]$, the *effective sample size* for arm k is defined as $n_k^{\text{eff}} := [\mathbb{E}\,[1/N_k(\mathcal{T})]]^{-1}$. Similarly, for any $r > 1$, the r-th order effective sample size is defined as $n_k^{\text{eff},r} := [\mathbb{E}\,[1/N_k^r(\mathcal{T})]]^{-1/r}$. Our next result exhibits a general risk bound on the Bregman risk that depends on the effective sample size.

Theorem 4.12. *Consider some adaptive sampling and stopping rules, and a fixed arm k. If there exists a constant $b \geq 1$ such that $N_k(\mathcal{T}) \geq b$ almost surely, then the risk of $\widehat{\mu}_k(\mathcal{T})$ is bounded as*

$$\mathbb{E}\left[D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}), \mu_k)\right] \leq \min\left\{2e\frac{1+\log(n_k^{\text{eff}}/b)}{n_k^{\text{eff}}}, \inf_{r>1}\frac{C_r}{n_k^{\text{eff},r}}\right\}, \tag{4.28}$$

where, for any $r > 1$,

$$C_r := \inf_{q\in(1,r)} \frac{2^{q/r}}{e}\frac{r^2}{(r-q)(q-1)}. \tag{4.29}$$

In particular, $C_r \to \infty$ as $r \to 1$.

Note that the bound in (4.28) is always non-negative since $n_k^{\text{eff}} \geq b$ by assumption. Further, if we always begin by sampling every arm once, then we may take $b = 1$ for $\mathcal{T} \geq K$. Of course, if we can choose a larger b, then the bound will be stronger. The proof of Theorem 4.12 can be found in Appendix C.3.2 and is based on the following deviation inequality for the unnormalized Bregman divergence loss, which is proved in Appendix C.3.1.

Lemma 4.13. *Under the assumptions in Theorem 4.12 we have that, for any $\delta \geq 0$,*

$$\mathbb{P}\left(D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}), \mu_k) \geq \delta\right) \leq 2\inf_{q\geq 1}\left[\mathbb{E}e^{-(q-1)\delta N_k(\mathcal{T})}\right]^{1/q} \leq 2e^{-\delta b}. \tag{4.30}$$

We remark that the results in Caballero et al. [1998], Peña et al. [2008] imply similar deviation inequalities and moment bounds for sub-Gaussian arms. The bound in Lemma 4.13 can be viewed as a generalization to sub-ψ arms.

We now convert the risk bound (4.28) into a bound on the expected ℓ_1 loss, and on the bias $|\widehat{\mu}_k(\mathcal{T}) - \mu_k|$. A minor complication arises due to the fact that the function ψ^* is strictly convex around 0 and, therefore, not invertible. Instead, we consider two invertible variants of ψ^*, both defined on $\Lambda^* \cap [0, \infty)$ and taking values in $[0, \infty)$:

$$z \mapsto \psi^*_+(z) = \psi^*(z) \quad \text{and} \quad z \mapsto \psi^*_-(z) = \psi^*(-z).$$

Corollary 4.14. *Suppose the assumptions in Theorem 4.12 hold. For each $k \in [K]$ and $b > 0$, define*

$$U_{k,b} := \min\left\{2e\frac{1+\log(n_k^{\text{eff}}/b)}{n_k^{\text{eff}}}, \inf_{r>1}\frac{C_r}{n_k^{\text{eff},r}}\right\}. \tag{4.31}$$

Then, the bias of the sample mean is bounded as

$$-\psi_-^{*\,-1}\left(U_{k,b}\right) \le \mathbb{E}\left[\widehat{\mu}_k(\mathcal{T})\right] - \mu_k \le \psi_+^{*\,-1}\left(U_{k,b}\right). \tag{4.32}$$

Furthermore, if ψ^* is symmetric around zero, then the ℓ_1 risk can be bounded as

$$\mathbb{E}\left|\widehat{\mu}_k(\mathcal{T}) - \mu_k\right| \le \psi_+^{*\,-1}\left(U_{k,b}\right). \tag{4.33}$$

The proof can be found in Appendix C.3.3. As one explicit example, if the underlying distribution is sub-Gaussian, $\psi_+^{*-1}(l) = \sigma\sqrt{2l}$ and the ℓ_1 risk of the sample mean is bounded as

$$\mathbb{E}\left|\widehat{\mu}_k(\mathcal{T}) - \mu_k\right| \le \sigma\sqrt{2U_{k,b}} = \sigma \min\left\{\sqrt{4e\frac{1+\log(n_k^{\text{eff}}/b)}{n_k^{\text{eff}}}}, \inf_{r>1}\sqrt{\frac{2C_r}{n_k^{\text{eff},r}}}\right\}. \tag{4.34}$$

We remark that the above bound on the bias is not improvable beyond the log factor in general by using the following stopped Brownian motion example [Siegmund, 1985, Ch. 3].

Example 4.15. *If we define a stopping time as the first time $W(t)$ exceeds a line with slope η and intercept $b > 0$, that is $\mathcal{T}_B := \inf\{t \ge 0 : W(t) \ge \eta t + b\}$, then for any slope $\eta \le \mu$, we have $\mathbb{E}\left[\frac{W(\mathcal{T}_B)}{\mathcal{T}_B} - \mu\right] = 1/b$.*

Note that a sum of Gaussians with mean μ behaves like a time-discretization of a Brownian motion with drift μ; since $\mathbb{E}W(t) = t\mu$, we may interpret $W(\mathcal{T}_B)/\mathcal{T}_B$ as a stopped sample mean, and the last equation implies that its bias is $1/b$ for any slope $\eta \le \mu$. In particular, if we set $\eta = \mu$, it is easy to deduce that $\mathbb{E}[1/\mathcal{T}_B] = 1/b^2$ and thus that

$$1/n^{\text{eff}} = \mathbb{E}\left[1/\mathcal{T}_B\right] = 1/b^2.$$

As a result, the bias of $W_{\mathcal{T}_B}/\mathcal{T}_B$ as a stopped sample mean is exactly equal to $\sqrt{1/n^{\text{eff}}}$ which matches (4.34) up to a log factor.

We end this section by discussing whether tight risk bounds can be obtained based on $\mathbb{E}N_k(\mathcal{T})$. By Jensen's inequality, it can be easily checked that $n_k^{\text{eff}} \le \mathbb{E}N_k(\mathcal{T})$. One may wonder if it is possible to obtain tighter bounds on both bias and risk that scale with $1/\mathbb{E}N_k(\mathcal{T})$ instead of n_k^{eff}. However, we can show that this is not possible in general. For instance, in the previous stopped Brownian motion case with $\eta = \mu$, we checked that the bias is equal to $1/b = 1/\sqrt{n^{\text{eff}}} > 0$. However, under the same setting, it is well-known that $\mathbb{E}N_k(\mathcal{T}) = \infty$. Therefore, the bias (namely $1/b$) can never be bounded by $1/\mathbb{E}N(\mathcal{T}) = 0$. Also, a risk bound in terms of $1/\mathbb{E}N_k(\mathcal{T})$ would imply consistency whenever $\mathbb{E}N_k(\mathcal{T}) \to \infty$, but Example 4.1 shows that $\widehat{\mu}_k$ can be inconsistent even when $\mathbb{E}N_k(\mathcal{T}) \to \infty$.

4.4.3 Bregman divergence risk bounds under fully adaptive settings

Let $I(\kappa; \mathcal{D}_\mathcal{T})$ be the mutual information between κ and the dataset $\mathcal{D}_\mathcal{T}$. When the dataset $\mathcal{D}_\mathcal{T}$ is collected in a deterministic manner, Russo and Zou [2016] showed how to bound the bias and expected ℓ_1 and ℓ_2 loss of adaptively chosen centered sub-Gaussian random variables by using $I(\kappa; \mathcal{D}_\mathcal{T})$. In particular, if each $\widehat{\mu}_k - \mu_k$ has mean zero and is $(\sigma/\sqrt{n})$-sub-Gaussian, then Russo and Zou [2016] proved that

$$\sqrt{n}\left|\mathbb{E}\widehat{\mu}_\kappa - \mu_\kappa\right| \le \sigma\sqrt{2I(\kappa; \mathcal{D}_\mathcal{T})}, \tag{4.35}$$

$$\mathbb{E}\left[\sqrt{n}\left|\widehat{\mu}_\kappa - \mu_\kappa\right|\right] \le \sigma\left(1 + c_1\sqrt{2I(\kappa; \mathcal{D}_\mathcal{T})}\right) \tag{4.36}$$

$$\mathbb{E}\left[n\left(\widehat{\mu}_\kappa - \mu_\kappa\right)^2\right] \le \sigma^2\left(1.25 + c_2 I(\kappa; \mathcal{D}_\mathcal{T})\right), \tag{4.37}$$

where $c_1 < 36$ and $c_2 \leq 10$ are universal constants.

For the reasons discussed in Section 4.3, however, these bounds are not directly applicable to the fully adaptive setting since each $\widehat{\mu}_k - \mu_k$ is no longer centered, due to the bias caused by adaptive sampling, stopping and rewinding. In particular, the bound for the bias given in equation (4.35) no longer holds under the fully adaptive setting because the bias can be non-zero even if κ is independent of $\mathcal{D}_{\mathcal{T}}$.

In this subsection we show that, by introducing an additional small "penalty for adaptivity", measured on the log-log scale, the bounds for deterministic and nonadpative sampling and stopping can be basically extended to the fully adaptive setting. Towards that end, and assuming that $N_k(t_0) > 3$ for all $k \in [K]$, we set

$$\widetilde{\widetilde{N}}_k(t) := \frac{N_k(t)}{\log\log N_k(t)}, \quad \forall k \in [K], \forall t \geq t_0. \tag{4.38}$$

We now present the main result of this section.

Theorem 4.16. *For any adaptive sampling and stopping rule and any adaptively chosen arm κ, suppose $\min_{k\in[K]} N_k(\tau) \geq b \geq 3$ almost surely for an adaptively rewound time $\tau \leq \mathcal{T}$. Then, the risk of $\widehat{\mu}_\kappa(\tau)$ is bounded as*

$$\mathbb{E}\left[\widetilde{\widetilde{N}}_\kappa(\tau) D_{\psi^*_{\mu_\kappa}}\left(\widehat{\mu}_\kappa(\tau), \mu_\kappa\right)\right] \leq C_b \left[I\left(\kappa; \mathcal{D}_{\mathcal{T}}\right) + 1.25\right], \tag{4.39}$$

where $C_b := 4e\left(1 + \frac{1}{\log\log b}\right)$.

Note that for the sub-Gaussian case, the inequality (4.39) is reduced to the following bound on the normalized ℓ_2 risk.

$$\mathbb{E}\left[\widetilde{\widetilde{N}}_\kappa(\tau)\left(\widehat{\mu}_\kappa(\tau) - \mu_\kappa\right)^2\right] \leq 2C_b\sigma^2\left[I\left(\kappa; \mathcal{D}_{\mathcal{T}}\right) + 1.25\right]. \tag{4.40}$$

By comparing the above bound with the bound (4.37) of Russo and Zou [2016], we can notice that our bound (4.39) under the fully adaptive setting only suffers a multiplicative normalization term which is of order $\log\log N_\kappa(\tau)$. Also, for a fixed target, the following example demonstrates that, in general, the bound (4.39) cannot be improved upon, aside from constants.

Example 4.17 (Example 4.9 revisited)**.** *In the same setting of Example 4.9, we further assume that each arm has a normal distribution variance σ_k^2. Then, from the definition of the stopping time $\mathcal{T}_b$ and the bound (4.39), we have following upper and lower bounds on the normalized ℓ_2 risk for a fixed target.*

$$\sigma_k^2 \leq \mathbb{E}\left[\widetilde{\widetilde{N}}_k(\mathcal{T}_b)\left(\widehat{\mu}_k(\mathcal{T}_b) - \mu_k\right)^2\right] \leq 2.5C_b\sigma_k^2, \tag{4.41}$$

in which upper and lower bounds are matched to each other up to a constant factor.

The proof of Theorem 4.16 in Appendix C.4.2 relies on the following deviation inequality for the normalized Bregman divergence loss, along with the Donsker-Varadhan variational representation of the KL divergence.

Lemma 4.18. *Consider some adaptive sampling and stopping rules. For a fixed $k \in [K]$ and a random time τ, assume $N_k(\tau) \geq b$ almost surely. Then, for any $\delta \geq 1$,*

$$\mathbb{P}\left(\widetilde{\widetilde{N}}_\kappa(\tau) D_{\psi^*_{\mu_k}}\left(\widehat{\mu}_k(\tau), \mu_k\right) \geq C_b\delta\right) \leq 2\exp\left\{-\delta\right\}. \tag{4.42}$$

The proof of the lemma is deferred to Appendix C.4.1. Similar inequalities have been developed in the context of always valid confidence sequences or finite-LIL bounds. Except in the sub-Gaussian case, the existing inequalities cannot be directly converted into bounds on the Bregman divergence. Recently, Garivier [2013] provided concentration inequalities for the KL loss, and Kaufmann and Koolen [2018] derived similar inequalities for the additive KL loss across several arms. However, their bounds depend on

δ in a complicated way making it difficult to develop bounds for the risk. In contrast, the bound in (4.42) is linear in δ which makes it easy to derive a bound of the risk in a simple form.

Next, for any $r > 0$, define the r-th order iterated logarithmically discounted effective sample size of an adaptively chosen arm as

$$\tilde{\tilde{n}}_{\kappa}^{\mathrm{eff},r} := \left[\mathbb{E}\left[1/\tilde{\tilde{N}}_{\kappa}^{r}(\tau)\right]\right]^{-1/r}, \quad \forall r > 0, \tag{4.43}$$

where the expectation is over the randomness in all four sources of adaptivity. This quantity is nonrandom, and the subscript κ merely differentiates it from the effective sample size of a fixed arm, and is not to be interpreted as residual randomness. We can also easily check that $\tilde{\tilde{n}}_{\kappa}^{\mathrm{eff},r}$ is decreasing with respect to r by using Jensen's inequality. The following corollary shows how to control risks of various orders by using $\tilde{\tilde{n}}_{\kappa}^{\mathrm{eff},r}$. The proof of the corollary can be found in Appendix C.4.3.

Corollary 4.19. *For any $r \in (0,1)$, the r-quasi-norm of the divergence can be bounded as*

$$\left[\mathbb{E} D_{\psi_{\mu_\kappa}^*}^{r}\left(\widehat{\mu}_\kappa(\tau), \mu_\kappa\right)\right]^{1/r} \leq \frac{C_b}{\tilde{\tilde{n}}_{\kappa}^{\mathrm{eff},r/(1-r)}} \left[I\left(\kappa; \mathcal{D}_{\mathcal{T}}\right) + 1.25\right]. \tag{4.44}$$

In the sub-Gaussian setting, by choosing $r = 1/2$, the above results immediately yields the bound for the ℓ_1 risk

$$\mathbb{E}\left|\widehat{\mu}_\kappa(\tau) - \mu_\kappa\right| \leq \frac{\sigma}{\sqrt{\tilde{\tilde{n}}^{\mathrm{eff}}}} \sqrt{2C_b\left[I\left(\kappa; \mathcal{D}_{\mathcal{T}}\right) + 1.25\right]}, \tag{4.45}$$

which is also comparable with the bound (4.36) on the ℓ_1 risk given by Russo and Zou [2016],

$$\mathbb{E}\left|\widehat{\mu}_\kappa(n) - \mu_\kappa\right| \leq \frac{\sigma}{\sqrt{n}} \left(c_1 \sqrt{2I(\kappa; \mathcal{D}_{\mathcal{T}})} + 1\right). \tag{4.46}$$

We quickly point out that the above theorem and corollary immediately yield results for the setting where we adaptively rewind to time τ, but choose a fixed arm $\kappa = k$, since $I(\kappa, \mathcal{D}_{\mathcal{T}}) = 0$ in this case. We also remark that by letting $r \to 1$, we get $\tilde{\tilde{n}}_{\kappa}^{\mathrm{eff},r/(1-r)} \to b/\log\log b$ which implies the following bound on the risk:

$$\mathbb{E} D_{\psi_{\mu_\kappa}^*}\left(\widehat{\mu}_\kappa(\tau), \mu_\kappa\right) \leq C_b \frac{\log\log b}{b} \left[I\left(\kappa; \mathcal{D}_{\mathcal{T}}\right) + 1.25\right]. \tag{4.47}$$

It is an open question whether it is possible to get a bound based on $\tilde{\tilde{n}}^{\mathrm{eff}}$ instead of b in the fully adaptive setting.

4.5 Summary of the main theorems and proof techniques

In this chapter, we have analyzed the behavior of the sample mean under four types of adaptivity implied by arbitrary rules for sampling, stopping, choosing and rewinding. Table 4.1 summarizes the risk bounds we have derived under different conditions on the distributions of the arms and under different data collection / analysis procedures.

The derivation of the upper bounds for the various notions of risks of the chosen means are based on the variational representations of the f_q-divergence (Theorem 4.6) and of the KL divergence (Theorem 4.16). These are given respectively by

$$\frac{1}{q} D_{f_q}(P||Q) = \sup_{f \in \mathcal{C}_p} \mathbb{E}_P\left[f(X)\right] - \mathbb{E}_Q\left[f(X)\right] - \mathbb{E}_Q\left[\frac{|f(X)|^p}{p}\right], \tag{4.48}$$

Table 4.1: Summary of normalized and unnormalized risk bounds under different conditions. (Recall that $\widetilde{N}_\kappa := N_\kappa / \log N_\kappa$ and $\widetilde{\widetilde{N}}_\kappa := N_\kappa / \log\log N_\kappa$.)

Tail condition	Data collection	Risk bound	
$\sigma_k^{(2)} < \infty$	Nonadaptive	$\mathbb{E}\left[N_k(T)\left(\widehat{\mu}_k(T)-\mu_k\right)^2\right] = \sigma_k^2$	(Prop 4.4)
$\max_k \sigma_k^{(2p)} < \infty$	Adaptive choosing	$\mathbb{E}\left[N_\kappa(T)\left(\widehat{\mu}_\kappa(T)-\mu_\kappa\right)^2\right] \le \|\sigma_\kappa\|_2^2 + C_p \left\|\sigma_\kappa^{(2p)}\right\|_{2p}^2 I_q^{1/q}(\kappa, \mathcal{D}_\mathcal{T})$	(Prop 4.5)
$\max_k \sigma_k^{(2(p+\epsilon))} < \infty$	Fully adaptive	$\mathbb{E}\left[\widetilde{N}_\kappa(\tau)\left(\widehat{\mu}_\kappa(\tau)-\mu_\kappa\right)^2\right] \le C_{1,\epsilon}\|\sigma_\kappa\|_2^2 + C_{p,\epsilon}\|\sigma_\kappa\|_{2p}^2 I_q^{1/q}(\kappa, \mathcal{D}_\mathcal{T})$	(Thm 4.6)
sub-ψ	Adaptive sampling and stopping	$\mathbb{E}\left[D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}), \mu_k)\right] \le \min\left\{2e^{\frac{1+\log(n_k^{\mathrm{eff}}/b)}{n_k^{\mathrm{eff}}}}, \inf_{r>1}\frac{C_r}{n_k^{\mathrm{eff},r}}\right\}$	(Thm 4.12)
sub-ψ	Fully adaptive	$\mathbb{E}\left[\widetilde{\widetilde{N}}_\kappa(\tau) D_{\psi^*_{\mu_\kappa}}(\widehat{\mu}_\kappa(\tau), \mu_\kappa)\right] \le C_b\left[I(\kappa; \mathcal{D}_\mathcal{T}) + 1.25\right]$	(Thm 4.16)

and

$$D_{KL}(P||Q) = \sup_{f \in \mathcal{C}_{\exp}} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q\left[e^{f(X)}\right], \tag{4.49}$$

where P, Q are probability measures on $\mathcal{X}$. In the first equation (4.48), $\mathcal{C}_p$ denotes the set of measurable functions $f\colon \mathcal{X} \mapsto \mathbb{R}$ such that $\mathbb{E}_Q|f(X)|^p < \infty$, with $p, q > 1$ satisfying $1/p + 1/q = 1$, while in the second equation (4.49), $\mathcal{C}_{\exp}$ is the set of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ such that $\mathbb{E}_Q\left[e^{f(X)}\right] < \infty$.

Now, for each $k \in [K]$, set $P_k = \mathcal{L}(\mathcal{D}_\mathcal{T} | \kappa = k)$, $Q = \mathcal{L}(\mathcal{D}_\mathcal{T})$ and

$$f_k = \begin{cases} \lambda \widetilde{N}_k(\tau)\left(\widehat{\mu}_k(\tau) - \mu_k\right)^2 & \text{(Theorem 4.6)}, \\ \lambda \widetilde{\widetilde{N}}_k(\tau) D_{\psi^*_{\mu_k}}(\widehat{\mu}_k, \mu_k) & \text{(Theorem 4.16)}. \end{cases} \tag{4.50}$$

By plugging these choices of P_k, Q and f_k in the right hand sides of (4.48) and (4.49), we obtain lower bounds on the f_q divergence and the KL divergence between the conditional and unconditional laws of the data. Next, based on these lower bounds, we derive upper bounds on the normalized risk of the chosen mean in Theorem 4.6 and 4.16. Detailed derivations can be found in Appendix C.2.2 (Theorem 4.6) and Appendix C.4.2 (Theorem 4.16).

This style of proof was originally developed by Russo and Zou [2016] and Jiao et al. [2017] for the fixed sample size setting. The main technical hurdle to extend it to the fully adaptive setting is to find tight upper bounds on the expectations of the pth power and the exponential moment of the normalized losses, defined for each $k \in [K]$ by

$$\begin{aligned} &\mathbb{E}\left[\left|\lambda \widetilde{N}_k(\tau)\left(\widehat{\mu}_k(\tau) - \mu_k\right)^2\right|^p\right] && \text{(Theorem 4.6)}, \\ &\mathbb{E}\left[\exp\left\{\lambda \widetilde{\widetilde{N}}_k(\tau) D_{\psi^*_{\mu_k}}(\widehat{\mu}_k, \mu_k)\right\}\right] && \text{(Theorem 4.16)}, \end{aligned}$$

where $\lambda > 0$ is a parameter to be chosen appropriately. To derive upper bounds independent of sampling and stopping rules, we use the deviation inequalities in Lemma 4.7 and 4.18 in conjunction with the following facts:

$$\begin{aligned} \mathbb{E}|X|^p &\le 1 + \int_1^\infty \mathbb{P}(|X| > \delta^{1/p}) \mathrm{d}\delta, \\ \mathbb{E}\left[e^X\right] &\le e + \int_1^\infty \mathbb{P}(X > \delta) e^\delta \mathrm{d}\delta. \end{aligned}$$

In the proofs of Lemma 4.7 and 4.18 we deploy martingale inequalities to obtain high probability bounds on events where the running sum $\{S_k(t)\}$ eventually exceeds certain linear functions of the number of draws $\{N_k(t)\}$. Specifically, in Lemma 4.7 we use the ℓ_p-version of the Dubins-Savage inequality [Khan, 2009], while in Lemma 4.18 our arguments are directly inspired by the proof of the exponential line-crossing inequality of [Howard et al., 2018a].

The derivation of the bound in Theorem 4.12 is based on the deviation inequality for unnormalized loss in Lemma 4.13 and the fact that $\mathbb{E}|X| = \epsilon + \int_\epsilon^\infty \mathbb{P}(|X| > \delta)\mathrm{d}\delta$ for any choice of $\epsilon \geq 0$; utilizing both, we have the following intermediate bound:

$$\mathbb{E}\left[D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}), \mu_k)\right] \leq \epsilon + 2\int_\epsilon^\infty \left[\mathbb{E}\exp\left\{-\frac{q}{p}\delta N_k(\mathcal{T})\right\}\right]^{1/q} \mathrm{d}\delta, \tag{4.51}$$

where $\epsilon \geq 0$ and $p, q > 1$ with $1/p + 1/q = 1$. By carefully choosing ϵ, p and q we then arrive at the final bounds in terms of effective sample sizes. The proof the deviation inequality in Lemma 4.13 is based on the following process:

$$\left\{\exp\left\{\lambda\left(S_k(t) - \mu_k N_k(t)\right) - N_k(t)\psi(\lambda)\right\}\right\},$$

which is martingale with respect to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$, for any fixed $\lambda \in \Lambda$.

4.6 Discussion and future work

We build on a line of interesting work that considered one type of adaptivity at a time. For example, the important work of Russo and Zou [2016] and its extensions by Jiao et al. [2017] can be viewed as understanding the bias of the sample mean under nonadaptive sampling, nonadaptive stopping and adaptive choosing. Similarly, the work by Nie et al. [2018] can be seen as providing a qualitative understanding of the sample mean under "optimistic" adaptive sampling, but for a deterministic arm stopped at a deterministic time. Further, while these past works have primarily focused on the bias, our work answers natural questions involving the estimation risk and consistency.

Several interesting questions remain fruitful for future research. The first one revolves around the choice of loss function for calculating the risk. Arguably, we picked the most natural loss function, which is the ℓ_2 loss for heavy-tailed arms and the Bregman divergence with respect to the convex conjugate of the CGF, also known as the KL-loss for exponential families. However, it is likely that the bounds achieved as implications of our results are not tight for other loss functions, and newer direct techniques may be more suitable. A second, related, question involves proving minimax lower bounds for risk (for various loss functions) under all kinds of adaptivity. The work of Sackrowitz and Samuel-Cahn [1986] on Bayes and minimax approaches towards evaluating a selected population may be a relevant starting point.

A final question revolves around possibly moving away from the sample mean, specifically whether there exist generic methods to either (a) alter the process of collecting the data to produce an unbiased estimator of the mean, or (b) to debias the sample mean posthoc given explicit knowledge of the exact sampling, stopping and choosing rule used. For aim (b), sample splitting was proposed by Xu et al. [2013], techniques from conditional inference were suggested by Nie et al. [2018], and a "one-step" estimator was suggested by Deshpande et al. [2018]. However, all three methods seemed to account for adaptive sampling, but not adaptive stopping or choosing, but their techniques seem to provide a good starting point. More recently, ideas from differential privacy were exploited by Neel and Roth [2018] for aim (a). It remains unclear what the theoretical and practical tradeoffs are between these methods, and how much they improve on the risk of the sample mean in a nonparametric and nonasymptotic sense in the fully adaptive setting.

Overall, we anticipate much progress on the above and other related questions in future years, due to the pressing concerns raised by the need to perform statistical inference on data collected via adaptive schemes that are common in the tech industry.

Appendix A

Appendix for Chapter 2

A.1 ϵ-greedy, UCB and Thompson sampling are optimistic sampling rules

A.1.1 Exploit and IIO conditions are sufficient for optimistic sampling

In Fact 2.7, we claimed that "Exploit" and "IIO" conditions in Nie et al. [2018] are jointly a sufficient condition for a sampling rule being optimistic. In this subsection, we formally restate Exploit and IIO conditions of Nie et al. [2018] in terms of our notations and prove Fact 2.7.

First, fix a deterministic stopping time T. Given any $t \in [T], k \in [K]$, define respectively the data from arm k until time t, and the data from all arms except k until time t, as

$$\mathcal{D}_t^{(k)} := \left\{X_{i,k}^*\right\}_{i=1}^{N_k(t)} \text{ and } \mathcal{D}_t^{(-k)} := \mathcal{D}_t \setminus \mathcal{D}_t^{(k)} = \bigcup_{j \neq k} \left\{X_{i,j}^*\right\}_{i=1}^{N_j(t)} \cup \{W_{-1}, W_0, \ldots, W_t\},$$

where $\mathcal{D}_t$ is the sample history up to time t under a tabular model $\mathcal{D}_\infty^*$. Let $\mathcal{D}_\infty^{*'}$ be another tabular model. Under $\mathcal{D}_\infty^{*'}$, we define $\mathcal{D}'_t, \mathcal{D}'^{(k)}_t$ and $\mathcal{D}'^{(-k)}_t$ in the same way. The Exploit condition in Nie et al. [2018] can be rewritten as following.

Definition A.1 (Exploit)**.** *Given any $t \in [T], k \in [K]$, suppose $\mathcal{D}_t^{(k)}$ and $\mathcal{D}'^{(k)}_t$ have the same size (that is $N'_k(t) = N_k(t)$) and $\mathcal{D}_t^{(-k)} = \mathcal{D}'^{(-k)}_t$. If the sample mean $\widehat{\mu}_k(t)$ under $\mathcal{D}_t^{(k)}$ is less than or equal to the sample mean $\widehat{\mu}'_k(t)$ under $\mathcal{D}'^{(k)}_t$, then*

$$\mathbb{1}(A_t = k) := f_{t,k}\left(\mathcal{D}_t^{(k)} \cup \mathcal{D}_t^{(-k)}\right) \leq f_{t,k}\left(\mathcal{D}'^{(k)}_t \cup \mathcal{D}_t^{(-k)}\right) =: \mathbb{1}(A'_t = k).$$

For the IIO condition, we present a specific version in the MAB setting which was originally used in Eq.(8) in the proof of Theorem 1 in Nie et al. [2018].

Definition A.2 (Independence of Irrelevant Options (IIO))**.** *For each t, k, the sampling random variable A_t can be written in terms of deterministic functions $f_{t,k}$ and $g_{t,k}$ such that*

$$A_t = \begin{cases} k & \text{if } f_{t,k}\left(\mathcal{D}_{t-1}\right) = 1 \\ j & \text{if } f_{t,k}\left(\mathcal{D}_{t-1}\right) = 0 \text{ and } g_{t,k}\left(\mathcal{D}_{t-1}^{(-k)}\right) = j \text{ for some } j \neq k. \end{cases}$$

Intuitively, $f_{t,k}$ is simply the indicator of whether arm k was pulled at time t; the crucial part is $g_{t,k}$, which specifies which arm is selected when arm k is not, and the IIO condition requires that $g_{t,k}$ ignores the data from arm k in order to determine which $j \neq k$ to pull instead.

It can be checked that ϵ-greedy, UCB and Thompson sampling under Gaussian arms and Gaussian priors satisfy both conditions. Indeed, if arm k is not the arm with the highest mean or highest UCB (for example), determining which other arm does get pulled in the next step does not depend on the data from arm k. In Appendix A.1.2, we present a sufficient condition for Thompson sampling to satisfy both conditions, and thus to be optimistic which shows Thompson sampling is optimistic for many commonly used exponential family arms including Gaussian, Bernoulli, exponential and Possion arms with their conjugate priors.

Before we prove Fact 2.7, we first introduce a lemma related to the IIO condition as follows.

Lemma A.1. *Fix a $k \in [K]$. Let $\mathcal{D}^*_\infty$ and $\mathcal{D}^{*'}_\infty$ be two MAB tabular representation that agree with each other except in their k-th column. Let $N_j(t)$ and $N'_j(t)$ be the numbers of draws from arm j for all $j \in [K]$ under $\mathcal{D}^*_\infty$ and $\mathcal{D}^{*'}_\infty$ respectively. Then, under IIO, the following implication holds:*

$$N_k(t) \le N'_k(t) \Rightarrow N_j(t) \ge N'_j(t), \quad \text{for all } j \neq k. \tag{A.1}$$

By switching the roles of $\mathcal{D}^*_\infty$ and $\mathcal{D}^{*'}_\infty$, we also have

$$N_k(t) \ge N'_k(t) \Rightarrow N_j(t) \le N'_j(t), \quad \text{for all } j \neq k, \tag{A.2}$$

and therefore,

$$N_k(t) = N'_k(t) \Rightarrow N_j(t) = N'_j(t), \quad \text{for all } j \neq k. \tag{A.3}$$

Proof of Lemma A.1. It is enough to prove the first statement. We follow the logic in the proof of Property 1 in Nie et al. [2018]. If $N_k(t) = t$ or $N'_k(t) = t$ then the claimed statement holds trivially since $N_j(t) + N_k(t) \le t$ and $N'_j(t) + N'_k(t) \le t$ for all $j \neq k$. Therefore, for the rest of the proof, we assume $N_k(t) \le N'_k(t) < t$.

For each t, define $s_1 < \cdots < s_{t-N_k(t)}$ to be the sequence of times at which arm k was *not* sampled before time t under $\mathcal{D}^*_\infty$. Similarly, let $s'_1 < \cdots < s'_{t-N_k(t)}$ be the sequence of times at which arm k was *not* sampled before time t under $\mathcal{D}^{*'}_\infty$. From the IIO condition and the assumption that $\mathcal{D}^*_\infty$ and $\mathcal{D}^{*'}_\infty$ agree with each other except in their k-th column, we have

$$A_{s_u} = A_{s'_u}, \quad \text{for all } u \in \{1, \dots, t - N'_k(t)\}, \tag{A.4}$$

which implies that

$$N'_j(t) = N'_j(s'_{t-N'_k(t)}) = N_j(s_{t-N'_k(t)}) \le N_j(s_{t-N_k(t)}) = N_j(t),$$

where the first and the last identities stem from the definition of s and s', the second identity is due to (A.4), and the inequality follows from the assumption that $N_k(t) \le N'_k(t)$ along with the fact that $u \mapsto s_u$ and $s \mapsto N_j(s)$ are increasing. □

Proof of Fact 2.7. Let us fix an arm k and a deterministic stopping time T, and a time $t \le T$, as required by Exploit and IIO conditions. The arguments below are inspired by case 1 in the proof of Theorem 1 in Nie et al. [2018].

Let $X^{*'}_{i,k}$ be an independent copy of $X^*_{i,k}$ and define $X^{*'}_\infty$ as a $\mathbb{N} \times K$ table which equals X^*_∞ on all entries except the (i,k)-th entry, which contains $X^{*'}_{i,k}$. Let $\mathcal{D}^{*'}_\infty = X^{*'}_\infty \cup \{W_{-1}, W_0, \dots\}$ denote the corresponding dataset, which only differs from $\mathcal{D}^*_\infty$ in one element. Let $N_k(T)$ and $N'_k(T)$ be numbers of draws from arm k up to time T based on $\mathcal{D}^*_\infty$ and $\mathcal{D}^{*'}_\infty$ respectively. Also for each $t \le T$, let A_t and A'_t be sampled arms based on $\mathcal{D}^*_\infty$ and $\mathcal{D}^{*'}_\infty$ respectively.

To prove the claim, it is enough to show that if $X^*_{i,k} \leq X^{*'}_{i,k}$ then $N_k(T) \leq N'_k(T)$ under Exploit and IIO conditions. Suppose, for the sake of deriving a contradiction, that there exist $i \in \mathbb{N}$ and $k \in [K]$ such that $X^*_{i,k} \leq X^{*'}_{i,k}$ but $N_k(T) > N'_k(T)$. Note that since A_s and A'_s are functions of the history up to time $s-1$, we know that $A_s = A'_s$ for all $s \leq t$, where t is defined as $t = \min\{s \geq 1 : N_k(s) = N'_k(s) = i\}$. If $t \geq T$, we have that $N_k(T) = N_k(t) = N'_k(t) = N'_k(T)$, which contradicts our assumption. Hence, we may assume $t < T$ for the rest of the proof.

Define $s_0 := \min\{s \geq 1 : N_k(s) > N'_k(s)\}$. From the definition of s_0, we know that $N_k(s_0 - 1) = N'_k(s_0 - 1)$. Since $\mathcal{D}^*_\infty$ and $\mathcal{D}^{*'}_\infty$ are identical except for their (i,k)-th entry, by Lemma A.1, we have that $N_j(s_0 - 1) = N'_j(s_0 - 1)$ for all j, which also implies that $\mathcal{D}_{s_0-1}$ and $\mathcal{D}'_{s_0-1}$ are identical except for the $N_k(t)$-th observation from arm k. Therefore, the sample mean from arm k up to time $s_0 - 1$ under $\mathcal{D}'_{s_0-1}$ is larger than the one under $\mathcal{D}_{s_0-1}$.

Then, by the Exploit condition, $A_{s_0} = k$ implies that $A'_{s_0} = k$. This contradicts the assumption that $N_k(s_0) > N'_k(s_0)$. Therefore, if $X^*_{i,k} \leq X^{*'}_{i,k}$ then $N_k(T)$ must be less than or equal to $N'_k(T)$. Since it holds for any $i \in \mathbb{N}$, $k \in [K]$ and T, the sampling strategy is optimistic, proving our claim that Exploit and IIO conditions are jointly a special case of an optimistic sampling rule. □

A.1.2 Sufficient conditions for Thompson sampling to be optimistic

In the previous subsection A.1.1, we show that Exploit and IIO conditions are jointly a sufficient condition for a sampling rule to be optimistic. In this subsection, we present a sufficient condition for Thompson sampling to satisfy both conditions, and thus to be optimistic.

For each k, let θ_k be the parameter of the distribution of arm k, and let $\mu_k = \mu(\theta_k)$. If we use an independent prior π on $\theta := (\theta_1, \ldots, \theta_K)$, it can be easily shown that posterior distributions of θ and $\mu(\theta) := (\mu(\theta_1), \ldots, \mu(\theta_K))$ are also coordinate-wise independent conditionally on the data. Therefore, the IIO condition is trivially satisfied for the Thompson sampling algorithms. However, it is difficult to check whether the Exploit condition is satisfied because there is no closed form for $\pi(k = \arg\max_{j \in [K]} \mu(\theta_j) | \mathcal{D}_t)$ in general.

One way to detour this issue is to study whether there exists a posterior sampling method such that the following statistically equivalent sampling algorithm satisfies the Exploit condition.

$$\nu_t(k) = \begin{cases} 1 & \text{if } k = \arg\max_{j \in [K]} \mu_j(\theta_{j,t-1}) \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta_{j,t-1}$ is a draw from the posterior distribution $\pi(\theta_j | \mathcal{D}_{t-1})$ at time $t-1$. If there exists such sampling method, we know that the sample mean from this Thompson sampling is negatively biased for any fixed k and T. With a slight abuse of notation, we say the Thompson sampling is optimistic in this case.

For example, in Appendix A.1.1, we show that Thompson sampling under Gaussian arm and Gaussian prior is optimistic by using a standard Gaussian posterior sampling method described in Section 1.1. Similarly, for the Bernoulli arm with parameters $\{p_k\}_{k=1}^K$ and beta prior with non-negative integer parameters (n, m) case, we can check that the corresponding Thompson sampling is optimistic using the equivalent optimistic sampling rule

$$\nu_t(k) = \begin{cases} 1 & \text{if } k = \arg\max_{j \in [K]} \frac{a_{j,t-1}}{a_{j,t-1} + b_{j,t-1}} \\ 0 & \text{otherwise,} \end{cases}$$

where $a_{j,t-1} = -\sum_{i=1}^{n+S_k(t-1)} \log U_{i,k}$, $b_{j,t-1} = -\sum_{i=1}^{m+N_k(t-1)-S_k(t-1)} \log W_{i,k}$ and each $U_{i,k}$ and $W_{i,k}$ are independent draws from $U(0,1)$.

In general, we have the following sufficient condition for the Thompson sampling to be optimistic.

Corollary A.2. *Suppose the distributions of the arms belong to a one-dimensional exponential family with density* $p_\eta(x) = \exp\{\eta T(x) - A(\eta)\}$ *with respect to some dominating measure* λ *and with* $\eta \in \mathrm{E}$. *Let* π *be a conjugate prior on* η *with a density proportional to* $\exp\{\tau\eta \ - n_0 A(\eta)\}$. *If* $\pi(\eta \leq x \mid \tau, n_0)$ *is a decreasing function of* τ *for any given* x *and* n_0, *and if* $\eta \mapsto \mu(\eta)$ *and* $x \mapsto T(x)$ *are both increasing or decreasing mappings, then Thompson sampling is optimistic.*

Proof. Fix a an arm $k \in [K]$. By the conjugacy, the posterior distribution for η_k given the data up to time t is given by

$$\pi\left(\eta_k | \mathcal{D}_t\right) \propto \exp\left\{\left(\tau + S_k^T(t)\right)\eta_k - \left(n_0 + N_k(t)\right) A(\eta_k)\right\},$$

where $S_k^T(t) := \sum_{s=1}^t \mathbb{1}(A_s = k) T(Y_s)$. Let $F\left(x | S_k^T(t), N_k(t)\right) := \pi\left(\eta_k \leq x | \mathcal{D}_t\right)$. From the condition on the prior, we know that $S_k^T(t) \ \mapsto \ F\left(x | S_k^T(t), N_k(t)\right)$ is a decreasing mapping for any given $x, N_k(t)$ and indices i, k and t. Therefore $S_k^T(t) \ \mapsto \ F^{-1}\left(y | S_k^T(t), N_k(t)\right)$ is an increasing mapping for any given $y, N_k(t)$ and indices i, k and t. Now, we can check that the Thompson sampling is equivalent to the following sampling rule.

$$\nu_t(k) = \begin{cases} 1 & \text{if } k = \arg\max_{j \in [K]} \mu\left(\eta_{j,t-1}\right) \\ 0 & \text{otherwise,} \end{cases}$$

where $\eta_{j,t-1} := F^{-1}\left(U_{j,t-1} | S_k^T(t-1), N_k(t-1)\right)$ and each $U_{j,t-1}$ is an independent draw from $U(0,1)$. Since $\eta \mapsto \mu(\eta)$ and $x \mapsto T(x)$ are both increasing (or decreasing), this sampling rule and the corresponding Thompson sampling is optimistic. □

We can check many commonly used one-dimensional exponential family arms with its conjugate prior satisfying the condition in Corollary A.2 which includes Gaussian distributions with a Gaussian prior, Bernoulli distributions with a beta prior, Poisson distributions with a gamma prior and exponential distributions with a gamma prior

A.1.3 Intuitions for the sign of the bias under each optimistic sampling and stopping

Under an optimistic sampling rule with a fixed stopping time and a fixed target, Xu et al. [2013] and Nie et al. [2018] provided some intuitions as to why the sample mean is negatively biased. In this subsection, we presents a similar intuitive explanation for the negative bias of the sample mean due to adaptive sampling. We also offer some intuition in order to explain the positive bias stemming from optimistic stopping rules in the one-armed case.

For an optimistic sampling rule with a fixed stopping time, assume for simplicity that we have a fixed target arm with a symmetric distribution around its true mean. Consider two equally possible realization of the experiment up to time t. In one realization, the sample mean at time t happens to be larger than its true mean. On the other hand, in the other scenario, the sample mean at time t happens to be smaller than its true mean. In the first case, the optimistic sampling rule will draw samples more often from the target arm, and thus the sample mean will regress more easily to its true mean. In contrast, in the other case, the optimistic sampling rule will draw samples less often and thus the sample mean is less likely to regress to its true mean due to the smaller sample size. Since these two realizations are equally likely, on average, the sample mean is negatively biased. See Figure A.1 for an illustration of this intuition.

For optimistic stopping in the one-armed case, consider the stopping rule that terminates the experiment when the sample mean crosses a predetermined upper boundary. See Figure A.2 for an illustrative stopping boundary. As we did for the sampling case, we again assume that the distribution of the arm is symmetric around its true mean. As before, consider two equally possible realizations. In one realization, the sample mean at early times happens to be larger than the true mean. On the other hand, in the other realization, the

sample means at early times is smaller than its true mean. In the first realization, the sample mean will cross the upper stopping boundary at an earlier time and thus the sample mean at the crossing time will be large. In contrast, in the other realization, the sample mean will cross the boundary at a later time and thus the optimistic stopping rule ensures that we will draw more samples in this realization and thus the sample mean is more likely to regress to its true mean due to the larger sample size. Since these two realizations are equally likely, on average, the sample mean is positively biased. See Figure A.2 for an illustration of this intuition.

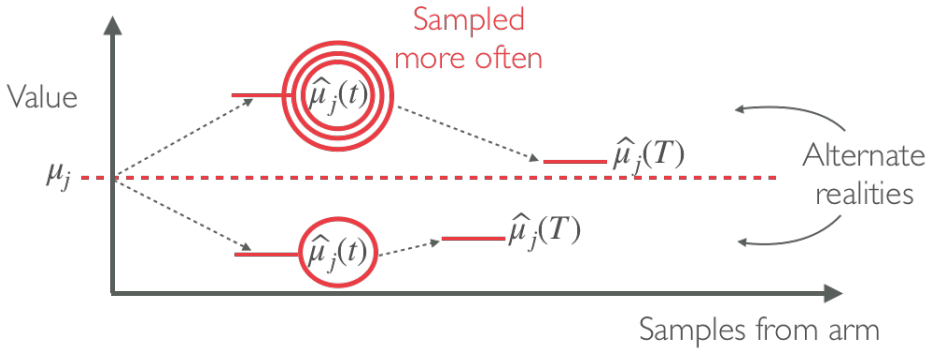

Figure A.1: *An illustration of the intuition for why optimistic sampling results in negative bias.*

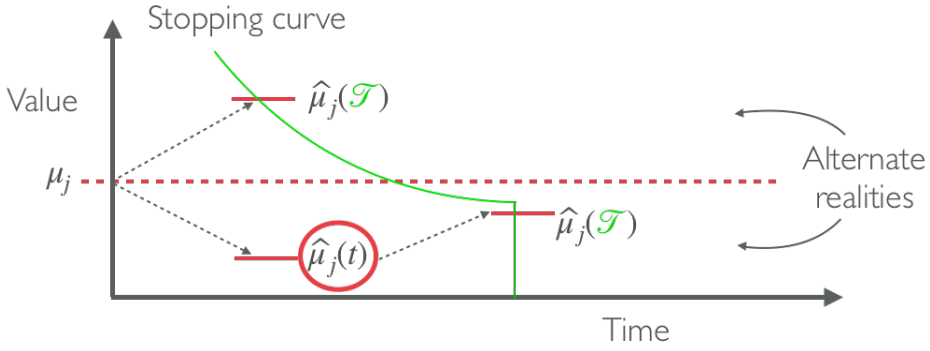

Figure A.2: *An illustration of the intuition for why optimistic stopping results in positive bias.*

A.2 Proofs

A.2.1 Proof of Theorem 2.10

Suppose that the data collecting strategy is monotonically increasing for the k-th distribution with $\mathbb{P}(\kappa = k) > 0$. From Theorem 3.1, it is sufficient to prove

$$\mathbb{E}\left[|\widehat{\mu}_k(\mathcal{T}) - \mu_k| \mid \kappa = k\right] < \infty \Longleftrightarrow \mathbb{E}\left[|\widehat{\mu}_k(\mathcal{T}) - \mu_k| \, \mathbb{1}\left(\kappa = k\right)\right] < \infty \tag{A.5}$$

From the strong law of large numbers, we know that $|\widehat{\mu}_k(\mathcal{T}) - \mu_k| \, \mathbb{1}(N_k(\mathcal{T}) = \infty) = 0$ almost surely. Hence, to prove Theorem 2.10, we can further refine the sufficient condition and it is enough to show

$$\mathbb{E}\left[|\widehat{\mu}_k(\mathcal{T}) - \mu_k| \, \mathbb{1}\left(\kappa = k\right) \mathbb{1}(N_k(\mathcal{T}) < \infty)\right] < \infty. \tag{A.6}$$

For each $t \geq 0$, define a process such that $L_0 = 0$ and

$$L_t := \sum_{s=1}^{t} \frac{\mathbb{1}(\kappa = k)}{N_k(\mathcal{T})} \mathbb{1}(N_k(\mathcal{T}) < \infty) \mathbb{1}\left(A_s = k\right) \left(Y_s - \mu_k\right), \quad \forall t \in \mathbb{N}. \tag{A.7}$$

Note that $L_t \to \left(\widehat{\mu}_k(\mathcal{T}) - \mu_k\right) \mathbb{1}\left(\kappa = k\right) \mathbb{1}(N_k(\mathcal{T}) < \infty)$ as $t \to \infty$ almost surely since it is understood that $\mathbb{1}(A_t = k) = 0$ for all $t > \mathcal{T}$. Therefore, to prove $\mathbb{E}\left[|\widehat{\mu}_k(\mathcal{T}) - \mu_k| \mathbb{1}\left(\kappa = k\right) \mathbb{1}(N_k(\mathcal{T}) < \infty)\right] < \infty$, it is enough show that there exists a random variable U with $\mathbb{E}[U] < \infty$ such that $|L_t| \leq U$ for all $t \geq 0$. Define U as

$$U = \sum_{s=1}^{\mathcal{T}} |L_s - L_{s-1}| = \sum_{s=1}^{\infty} |L_s - L_{s-1}| \mathbb{1}\left(\mathcal{T} \geq s\right). \tag{A.8}$$

Clearly, $|L_t| \leq U$ for all t. In order to show that $\mathbb{E}[U] < \infty$, first note that for any $t \geq 1$, we have

$$\begin{aligned}
\mathbb{E}\left[|L_{t+1} - L_t| \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\frac{\mathbb{1}\left(\kappa = k\right)}{N_k(\mathcal{T})} \mathbb{1}(N_k(\mathcal{T}) < \infty) \mathbb{1}\left(A_{t+1} = k\right) |Y_{t+1} - \mu_k| \mid \mathcal{F}_t\right] \\
&\leq \mathbb{E}\left[\mathbb{1}(A_{t+1} = k) |Y_{t+1} - \mu_k| \mid \mathcal{F}_t\right] \\
&= \mathbb{1}(A_{t+1} = k) \mathbb{E}\left[|Y_{t+1} - \mu_k| \mid \mathcal{F}_t\right] \\
&= \mathbb{1}(A_{t+1} = k) \int |x - \mu_k| \mathrm{d}P_k(x) \\
&:= c_k \mathbb{1}(A_{t+1} = k),
\end{aligned} \tag{A.9}$$

where the first inequality comes from the assumption $N_k(\mathcal{T}) \geq 1$ for all k with $\mathbb{P}(\kappa = k) > 0$, and the following equality holds because $\mathbb{1}(A_{t+1} = k) \in \mathcal{F}_t$. The third equality stems from the observation that, on the event $(A_{t+1} = k)$, $Y_{t+1} \sim P_k$ and it is independent of the previous history. Therefore, we obtain that

$$\begin{aligned}
\mathbb{E}[U] &= \sum_{s=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[|L_s - L_{s-1}| \mathbb{1}\left(\mathcal{T} \geq s\right) \mid \mathcal{F}_{s-1}\right]\right] \\
&= \sum_{s=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left(\mathcal{T} \geq s\right) \mathbb{E}\left[|L_s - L_{s-1}| \mid \mathcal{F}_{s-1}\right]\right] \quad (\text{since } \mathbb{1}\left(\mathcal{T} \geq s\right) \in \mathcal{F}_{s-1}.) \\
&\leq c_k \sum_{s=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left(A_s = k\right) \mathbb{1}\left(\mathcal{T} \geq s\right)\right] \quad (\text{by the inequality (A.9)}) \\
&= c_k \mathbb{E} N_k(\mathcal{T}) < \infty,
\end{aligned}$$

where the finiteness of the last term follows from the assumption $\mathbb{E}N_k(\mathcal{T}) < \infty$ for all k with $\mathbb{P}(\kappa = k) > 0$. This proves the inequality (2.2). The next inequality (2.3) follows immediately from this result and the identity

$$\mathbb{E}\left[\widehat{\mu}_\kappa(\mathcal{T}) - \mu_\kappa\right] = \sum_{k:\mathbb{P}(\kappa=k)>0} \mathbb{E}\left[\widehat{\mu}_\kappa(\mathcal{T}) - \mu_\kappa \mid \kappa = k\right] \mathbb{P}(\kappa = k).$$

Thus, the sample mean at the stopping time $\mathcal{T}$ is negatively biased.

If the data collecting strategy is monotonically increasing, the supermartingale is replaced by a submartingale and the inequalities are reversed. This observation completes the proof.

Now, suppose each arm has a bounded distribution. without loss of generality, assume there exists a fixed $M > 0$ such that $P_k([\mu_k - M, \mu_k + M]) = 1$ for all $k \in [K]$. Then for any $t \geq 1$, we have

$$\begin{aligned}\mathbb{E}\left[|L_{t+1} - L_t| \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\frac{\mathbb{1}(\kappa = k)}{N_k(\mathcal{T})}\mathbb{1}(N_k(\mathcal{T}) < \infty)\,\mathbb{1}(A_{t+1} = k)\,|Y_{t+1} - \mu_k| \mid \mathcal{F}_t\right] \\ &\leq M\mathbb{E}\left[\frac{\mathbb{1}(A_{t+1} = k)}{N_k(\mathcal{T})}\mathbb{1}(N_k(\mathcal{T}) < \infty) \mid \mathcal{F}_t\right].\end{aligned} \tag{A.10}$$

Therefore, we obtain that

$$\begin{aligned}\mathbb{E}[U] &= \sum_{s=1}^{\infty}\mathbb{E}\left[\mathbb{E}\left[|L_s - L_{s-1}|\,\mathbb{1}(\mathcal{T} \geq s) \mid \mathcal{F}_{s-1}\right]\right] \\ &= \sum_{s=1}^{\infty}\mathbb{E}\left[\mathbb{1}(\mathcal{T} \geq s)\,\mathbb{E}\left[|L_s - L_{s-1}| \mid \mathcal{F}_{s-1}\right]\right] \quad (\text{since } \mathbb{1}(\mathcal{T} \geq s) \in \mathcal{F}_{s-1}) \\ &\leq M\sum_{s=1}^{\infty}\mathbb{E}\left[\frac{\mathbb{1}(A_s = k)}{N_k(\mathcal{T})}\mathbb{1}(N_k(\mathcal{T}) < \infty)\,\mathbb{1}(\mathcal{T} \geq s)\right] \quad (\text{by the inequality (A.10)}) \\ &= M < \infty \quad (\text{by the definition of } N_k(\mathcal{T})),\end{aligned}$$

which implies that if each arm has a bounded distribution, we can determine the sign of the bias of the sample mean at the stopping time $\mathcal{T}$ without assuming $\mathbb{E}N_k(\mathcal{T}) < \infty$ for all k with $\mathbb{P}(\kappa = k) > 0$.

About Remark 2.11. In our recent work [Shin et al., 2019a], we showed that if arm k has a finite p-th moment for a fixed $p > 2$, the following bound on the normalized ℓ_2 risk of the sample mean holds:

$$\mathbb{E}\left[\frac{N_k(\mathcal{T})}{\log N_k(\mathcal{T})}\left(\widehat{\mu}_k(\mathcal{T}) - \mu_k\right)^2\right] < \infty, \tag{A.11}$$

provided that $N_k(\mathcal{T}) \geq 3$. In this case, we can show that $\mathbb{E}[U] < \infty$ without assuming $\mathbb{E}N_k(\mathcal{T}) < \infty$, where U is defined in (A.8). For each k, set $c_k := \int |x - \mu_k| \mathrm{d}P_k(x)$. Let $\widehat{c}_k(\mathcal{T})$ be the sample mean estimator of c_k at the stopping time $\mathcal{T}$. Then, we have

$$\begin{aligned}\mathbb{E}[U] &= \sum_{s=1}^{\infty}\mathbb{E}\left[\mathbb{E}\left[|L_s - L_{s-1}|\,\mathbb{1}(\mathcal{T} \geq s) \mid \mathcal{F}_{s-1}\right]\right] \\ &= \sum_{s=1}^{\infty}\mathbb{E}\left[\frac{\mathbb{1}(\kappa = k)}{N_k(\mathcal{T})}\mathbb{1}(A_s = k)\,|Y_s - \mu_k|\,\mathbb{1}(\mathcal{T} \geq s)\right] \\ &\leq \mathbb{E}\left[\sum_{s=1}^{\infty}\frac{\mathbb{1}(A_s = k)}{N_k(\mathcal{T})}\,|Y_s - \mu_k|\,\mathbb{1}(\mathcal{T} \geq s)\right] \\ &:= \mathbb{E}\left[\widehat{c}_k(\mathcal{T})\right] \\ &\leq \mathbb{E}\,|\widehat{c}_k(\mathcal{T}) - c_k| + c_k \\ &\leq \mathbb{E}\left[\sqrt{\frac{N_k(\mathcal{T})}{\log N_k(\mathcal{T})}}\,|\widehat{c}_k(\mathcal{T}) - c_k|\right] + c_k \\ &\leq \sqrt{\mathbb{E}\left[\frac{N_k(\mathcal{T})}{\log N_k(\mathcal{T})}\left(\widehat{c}_k(\mathcal{T}) - c_k\right)^2\right]} + c_k < \infty,\end{aligned}$$

where in the last bound we have used (A.11). Thus, if each arm has a finite p-th moment for a fixed $p > 2$, we can determine the sign of the bias of the sample mean at the stopping time $\mathcal{T}$ without assuming $\mathbb{E}N_k(\mathcal{T}) < \infty$ for all k with $\mathbb{P}(\kappa = k) > 0$.

A.2.2 Proof of Corollary 2.12 (The lil'UCB algorithm results in positive bias)

Before presenting a formal proof of Corollary 2.12, we first provide an intuitive explanation why any reasonable and efficient algorithm for the best-arm identification problem would result in positive bias. For any $k \in [K]$ and $i \in \mathbb{N}$, let $\mathcal{D}^*_\infty$ and $\mathcal{D}^{*'}_\infty$ be two MAB tabular representation that agree with each other except $X^*_{i,k} < X^{*'}_{i,k}$. Since we have a larger value from arm k in the second scenario $\mathcal{D}^{*'}_\infty$, if $\kappa = k$ under the first scenario $\mathcal{D}^*_\infty$, any reasonable algorithm would also pick the arm k under the more favorable scenario $\mathcal{D}^{*'}_\infty$. In this case, we know that $\kappa = k$ implies $\kappa' = k$. Also note that any efficient algorithm should be able to exploit the more favorable scenario $D^{*'}_\infty$ to declare arm k as the best arm by using less samples from arm k. Therefore, we would have $N_k(\mathcal{T}) \geq N'_k(\mathcal{T}')$. In sum, we can expect that, from any reasonable and efficient algorithm, we would have $\frac{\mathbb{1}(\kappa=k)}{N_k(\mathcal{T})} \leq \frac{\mathbb{1}(\kappa'=k)}{N'_k(\mathcal{T}')}$ which shows that the algorithm would be monotonically increasing and thus the sample mean of the chosen arm is positively biased. Below, we formally verify that this intuition works for the lil'UCB algorithm.

Proof of Corollary 2.12. For any given i, k, let $X^{*'}_{i,k}$ be an independent copy of $X^*_{i,k}$ and define $X^{*'}_\infty$ as a $\mathbb{N} \times K$ table which equals X^*_∞ on all entries except the (i,k)-th entry, which contains $X^{*'}_{i,k}$. Let $\mathcal{D}^{*'}_\infty = X^{*'}_\infty \cup \{W_{-1}, W_0, \dots\}$ denote the corresponding dataset, which only differs from $\mathcal{D}^*_\infty$ in one element. Let $(N_k(T), N'_k(T))$ denote the numbers of draws from arm k up to time T. Let $(\mathcal{T}, \mathcal{T}')$ be the stopping times and (κ, κ') be choosing functions as determined by the lil'UCB algorithm under $\mathcal{D}^*_\infty$ and $\mathcal{D}^{*'}_\infty$ respectively.

Suppose $X^*_{i,k} \leq X^{*'}_{i,k}$. Proving that the lil'UCB algorithm is monotonically increasing (and hence results in positive bias) corresponds to showing that the following inequality holds:

$$\frac{\mathbb{1}(\kappa = k)}{N_k(\mathcal{T})} \leq \frac{\mathbb{1}(\kappa' = k)}{N'_k(\mathcal{T}')}. \tag{A.12}$$

If $\kappa \neq k$, the inequality (A.12) holds trivially. Therefore, for the rest of the proof, we assume $\kappa = k$ which also implies $\mathcal{T} < \infty$. (If not, the lil'UCB algorithm is not stopped, and thus $\kappa \neq k$.)

First, we can check that the lil'UCB sampling is a special case of UCB-type sampling algorithms. Therefore, it is an optimistic sampling method which implies that for any *fixed* $t > 0$, and *fixed* arm k, we have $N_k(t) \leq N'_k(t)$. Since $\sum_{j \neq k} N_j(t) = t - N_k(t)$ for all t, we can rewrite the lil'UCB stopping rule as stopping the sampling whenever there exists a k such that N_k, which is a non-decreasing function of t, crosses the strictly increasing linear boundary $\left\{(n,t) : n = \frac{1+\lambda t}{1+\lambda}\right\}$ for a fixed $\lambda > 0$. Since $N_k(t) \leq N'_k(t)$ for all t, we know that $\mathcal{T}' \leq \mathcal{T}$.

Since the linear boundary is increasing, we can check $N'_k(\mathcal{T}') \leq N_k(\mathcal{T})$ if $\kappa' = k$. Therefore, to complete the proof, it is enough to show that $\kappa = k$ implies $\kappa' = k$. For the sake of deriving a contradiction, assume $\kappa = k$ but $\kappa' \neq k$. Then, there exists $j \neq k$ such that $\kappa' = j$. By the definition of κ', it is equivalent to $N'_j(\mathcal{T}') = \max_{l \in [K]} N'_l(\mathcal{T}')$. Hence, we have that

$$N'_j(\mathcal{T}') > N'_k(\mathcal{T}'). \tag{A.13}$$

Similarly, we can show that

$$N_j(\mathcal{T}) < N_k(\mathcal{T}). \tag{A.14}$$

Since $\mathcal{T}'$ is the first time t such that, for some l, $N'_l(t)$ has crossed the boundary, we know that j is also the index of the arm which has crossed the boundary first time. Also, since the lil'UCB sampling satisfies the IIO condition, Lemma A.1 along with the fact that $N_k(t) \leq N'_k(t)$ for all t implies that $N_j(t) \geq N'_j(t)$ for all $j \neq k$. From the two observations above, we have the following inequalities:

$$\frac{1+\lambda\mathcal{T}'}{1+\lambda} \leq N'_j(\mathcal{T}') \leq N_j(\mathcal{T}'),$$

which implies that $t \mapsto N_j(t)$ is crossing the boundary at time $\mathcal{T}'$. By the definition of $\mathcal{T}$ and, by assumption, $\kappa = k$, we obtain that $\mathcal{T} \leq \mathcal{T}'$.

Similarly, from the fact that $N_k(t) \leq N'_k(t)$ for all t along with the definition of $\mathcal{T}$, we have that

$$\frac{1+\lambda\mathcal{T}}{1+\lambda} \leq N_k(\mathcal{T}) \leq N'_k(\mathcal{T}),$$

which implies that $t \mapsto N'_k(t)$ is crossing the boundary at time $\mathcal{T}$, and thus $\mathcal{T}' \leq \mathcal{T}$ since $\kappa' \neq k$ by assumption.

From the two observations above, we have $\mathcal{T}' = \mathcal{T}$. Finally, note that

$$N'_k(\mathcal{T}') < N'_j(\mathcal{T}') \leq N_j(\mathcal{T}') = N_j(\mathcal{T}) < N_k(\mathcal{T}) \leq N'_k(\mathcal{T}) = N'_k(\mathcal{T}')$$

where the first inequality comes from the inequality (A.13). The second inequality come from $N'_j \leq N_j$. The first equality comes from $\mathcal{T}' = \mathcal{T}$ and the third inequality comes from the inequality (A.14). The last inequality comes from $N_k \leq N'_k$ and the final equality comes from $\mathcal{T} = \mathcal{T}'$.

This is a contradiction, and, therefore, $\kappa = k$ implies that $\kappa' = k$. This proves that the lil'UCB algorithm is monotonically increasing and the chosen stopped sample mean from the lil'UCB algorithm is positively biased. □

A.2.3 Proof of Proposition 2.8 (bias expression) via Lemma 2.9 (Wald's identity for MAB)

By direct substitution, we first note that

$$\begin{aligned}
\mathbb{E}\left|S_k(\mathcal{T}) - \mu_k N_k(\mathcal{T})\right| &= \mathbb{E}\left[\sum_{t=1}^{\infty} \mathbb{1}\left(A_t = k\right)\left|Y_t - \mu_k\right| \mathbb{1}\left(\mathcal{T} \geq t\right)\right] \\
&= \sum_{t=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left(A_t = k\right)\left|Y_t - \mu_k\right| \mathbb{1}\left(\mathcal{T} \geq t\right)\right] \\
&= \sum_{t=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left(A_t = k\right) \mathbb{1}\left(\mathcal{T} \geq t\right) \mathbb{E}\left[\left|Y_t - \mu_k\right| \mid \mathcal{F}_{t-1}\right]\right] \\
&= \sum_{t=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left(A_t = k\right) \mathbb{1}\left(\mathcal{T} \geq t\right) \int \left|x - \mu_k\right| \mathrm{d}P_k(x)\right] \\
&= \int \left|x - \mu_k\right| \mathrm{d}P_k(x) \mathbb{E}\left[\sum_{t=1}^{\infty} \mathbb{1}\left(A_t = k\right) \mathbb{1}\left(\mathcal{T} \geq t\right)\right] \\
&= \int \left|x - \mu_k\right| \mathrm{d}P_k(x) \mathbb{E}\left[N_k(\mathcal{T})\right] < \infty,
\end{aligned}$$

where the second equality comes from the Tonelli's theorem and the third equality stems from the facts that $\mathbb{1}(A_t = k)$ and $\mathbb{1}(\mathcal{T} \geq t)$ are $\mathcal{F}_{t-1}$ measurable. The fourth equality comes from the fact that, on

event $\mathbb{1}(A_t = k)$, $Y_t \sim P_k$ and it is independent of the previous history. Finally, the finiteness of the last term comes from the assumption of the existence of the first moment of k-th arm and $\mathbb{E}[N_k(\mathcal{T})] < \infty$. Therefore, by the dominated convergence theorem, we have

$$\begin{aligned}
\mathbb{E}\left[S_k(\mathcal{T}) - \mu_k N_k(\mathcal{T})\right] &= \mathbb{E}\left[\sum_{t=1}^{\infty} \mathbb{1}\left(A_t = k\right)\left[Y_t - \mu_k\right]\mathbb{1}\left(\mathcal{T} \geq t\right)\right] \\
&= \sum_{t=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left(A_t = k\right)\left[Y_t - \mu_k\right]\mathbb{1}\left(\mathcal{T} \geq t\right)\right] \\
&= \sum_{t=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left(A_t = k\right)\mathbb{1}\left(\mathcal{T} \geq t\right)\mathbb{E}\left[Y_t - \mu_k \mid \mathcal{F}_{t-1}\right]\right] \\
&= 0,
\end{aligned}$$

which implies $\mu_k \mathbb{E}\left[N_k(\mathcal{T})\right] = \mathbb{E}\left[S_k(\mathcal{T})\right]$, which proves the generalization of Wald's first identity.

Since $\mathbb{E}\left[N_k(\mathcal{T})\right] > 0$, one can then express μ_k as

$$\mu_k = \frac{\mathbb{E}\left[S_k(\mathcal{T})\right]}{\mathbb{E}[N_k(\mathcal{T})]}.$$

By direct substitution, the bias of the sample mean can thus be expressed as

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mu}_k(\mathcal{T}) - \mu_k\right] &= \mathbb{E}\left[\widehat{\mu}_k(\mathcal{T})\left(1 - \frac{N_k(\mathcal{T})}{\mathbb{E}[N_k(\mathcal{T})]}\right)\right] \\
&= \operatorname{Cov}\left(\widehat{\mu}_k(\mathcal{T}), \left(1 - \frac{N_k(\mathcal{T})}{\mathbb{E}[N_k(\mathcal{T})]}\right)\right) \\
&= -\frac{\operatorname{Cov}\left(\widehat{\mu}_k(\mathcal{T}), N_k(\mathcal{T})\right)}{\mathbb{E}[N_k(\mathcal{T})]}.
\end{aligned}$$

This completes the proof of the proposition.

A.3 Additional simulation results

A.3.1 More on negative bias due to optimistic sampling

We conduct a simulation study in which we have three unit-variance Gaussian arms with $\mu_1 = 1, \mu_2 = 2$ and $\mu_3 = 3$. After sampling once from each arm, greedy, UCB and Thompson sampling are used to continue sampling until $T = 200$. We repeat the whole process from scratch 10^4 times for each algorithm to get an accurate estimate for the bias.

For UCB, we use $u_{t-1}(s, n) = \sqrt{\frac{2\log(1/\delta)}{n}}$ with $\delta = 0.1$. For Thompson sampling, we use independent standard Normal priors for simplicity. We repeat the whole process from scratch 2000 times for each algorithm to get an accurate estimate for the bias.

Figure A.3 shows the distribution of observed differences between sample means and the true mean for each arm under the greedy algorithm. Vertical lines correspond to biases. The example demonstrates that the sample mean is negatively biased under optimistic sampling rules. Similar results from UCB / Thompson sampling algorithms can be found in Section 2.2.1.

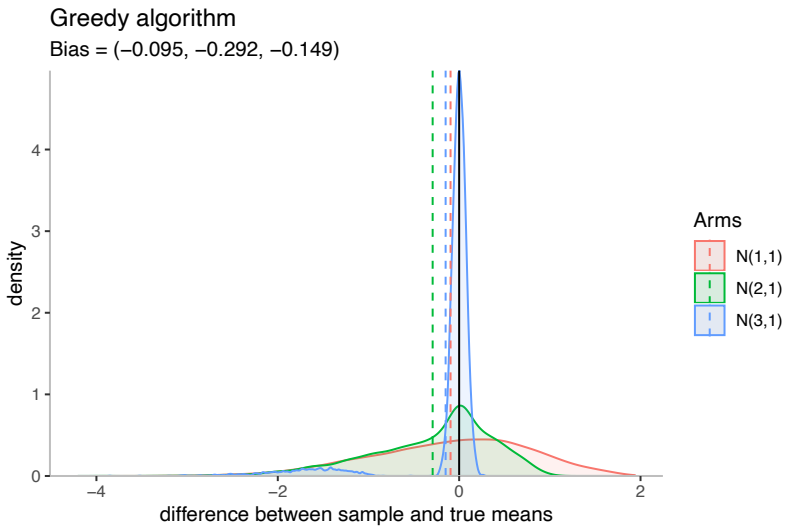

Figure A.3: *Data is collected by the greedy algorithm from three unit-variance Gaussian arms with* $\mu_1 = 1, \mu_2 = 2$ *and* $\mu_3 = 3$*. For all three arms, sample means are negatively biased.*

A.3.2 Positive bias from optimistic choosing and stopping in identifying the largest mean

Suppose we have K arms with mean $\mu_1, \ldots, \mu_K$. As we were in Section 2.2.3, we are interested not in each individual arm but in the arm with the largest mean. That is, our target of inference is $\mu_* := \max_{k \in [K]} \mu_k$.

Instead of using the lil'UCB algorithm, we can draw a sample from each arm in a cyclic order for each time t and use a naive sequential procedure based on the following stopping time.

$$\mathcal{T}_M^\delta := \inf \left\{ t \in \{K, 2K, \ldots, MK\} : \widehat{\mu}_{(1)}(t) > \widehat{\mu}_{(2)}(t) + \delta \right\}, \tag{A.15}$$

where $M, \delta > 0$ are prespecified constants and $\widehat{\mu}_{(k)}(t)$ is the k-th largest sample mean at time t. Once we stop sampling at time $\mathcal{T}_M^\delta$, we can estimate the largest mean by the largest stopped sample mean $\widehat{\mu}_{(1)}\left(\mathcal{T}_M^\delta\right)$.

The performance of this sequential procedure can vary based on underlying distribution of the arm and the choice of δ and M. However, we can check this optimistic choosing and stopping rules are jointly monotonic increasing and thus the largest stopped sample mean $\widehat{\mu}_{(1)}\left(\mathcal{T}_M^\delta\right)$ is always positively based for any choice of δ and M.

To verify it with a simulation, we set 3 unit-variance Gaussian arms with means $(\mu_1, \mu_2, \mu_3) = (g, 0, -g)$ for each gap parameter $g = 1, 3, 5$. We conduct 10^4 trials of this sequential procedure with $M = 1000$ and $\delta = 0.7 \times g$. Figure A.4 shows the distribution of observed differences between the chosen sample means and the corresponding true mean for each δ. Vertical lines correspond to biases. The simulation study demonstrate that, in all configurations, the largest stopped sample mean $\widehat{\mu}_{(1)}\left(\mathcal{T}_M^\delta\right)$ is always positively biased. Note, in contrast to the lil'UCB case in Section 2.2.3, we have a larger bias for a smaller gap since the number of sample sizes are similar for each gaps due to the adaptive (and oracle) choice of the parameter δ but a smaller gap makes more difficult to identify largest mean correctly.

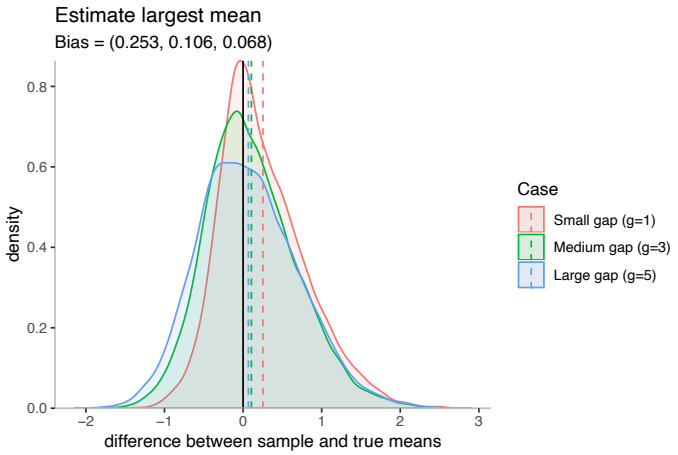

Figure A.4: *Data is collected by the sequential procedure described in Appendix A.3.2 under unit-variance Gaussian arms with* $\mu_1 = g, \mu_2 = 0$ *and* $\mu_3 = -g$ *for each gap parameter* $g = \{1, 3, 5\}$*. For each gap* g*, we set the parameter* $\delta = 0.7 \times g$ *and* $M = 1000$*. For all cases, chosen sample means are positively biased.*

Appendix B

Appendix for Chapter 3

B.1 Proofs

B.1.1 Proof of Theorem 3.1

Under the condition in Theorem 3.1, we first prove that if the function $\mathcal{D}_\infty^* \mapsto \mathbb{1}(C)/N_k(\mathcal{T})$ is an decreasing function of $X_{i,k}^*$ while keeping all other entries in $\mathcal{D}_\infty^*$ fixed for each i then, for any $t \in \mathbb{N}$ and $y \in \mathbb{R}$, the following inequality holds.

$$\mathbb{E}\left[\frac{\mathbb{1}(C)}{N_k(\mathcal{T})}\mathbb{1}(N_k(\mathcal{T}) < \infty)\mathbb{1}(A_t = k)\left[\mathbb{1}(Y_t \le y) - F_k(y)\right]\right] \ge 0 \tag{B.1}$$

Proof of inequality (B.1). First note that if $\mathcal{D}_\infty^* \mapsto \mathbb{1}(C)/N_k(\mathcal{T})$ is a decreasing function of $X_{i,k}^*$ then the following function is also a decreasing function of $X_{i,k}^*$:

$$D_\infty^* \mapsto \frac{\mathbb{1}(C)}{N_k(\mathcal{T})}\mathbb{1}(N_k(\mathcal{T}) < \infty) := h(\mathcal{D}_\infty^*) \tag{B.2}$$

Then, by the tabular representation of MAB, we can rewrite the LHS of inequality (B.1) as follows:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\mathbb{1}(C)}{N_k(\mathcal{T})}\mathbb{1}(N_k(\mathcal{T}) < \infty)\mathbb{1}(A_t = k)\left[\mathbb{1}(Y_t \le y) - F_k(y)\right]\right] \\
&= \mathbb{E}\left[h(\mathcal{D}_\infty^*)\mathbb{1}(A_t = k)\left[\mathbb{1}(X_{N_k(t),k}^* \le y) - F_k(y)\right]\right] \\
&= \mathbb{E}\sum_{i=1}^{t}\left[h(\mathcal{D}_\infty^*)\mathbb{1}(A_t = k, N_k(t) = i)\left[\mathbb{1}(X_{i,k}^* \le y) - F_k(y)\right]\right] \\
&= \sum_{i=1}^{t}\mathbb{E}\left[h(\mathcal{D}_\infty^*)\mathbb{1}(A_t = k, N_k(t) = i)\left[\mathbb{1}(X_{i,k}^* \le y) - F_k(y)\right]\right],
\end{aligned}$$

where the third equality comes from the fact $N_k(t) \in \{1, \ldots, t\}$. Therefore, to prove the inequality (B.1), it is enough to show the following inequality:

$$\mathbb{E}\left[\mathbb{1}(A_t = k, N_k(t) = i)\, h(\mathcal{D}_\infty^*)\left[\mathbb{1}(X_{i,k}^* \le y) - F_k(y)\right]\right] \ge 0. \tag{B.3}$$

Note that the term $\mathbb{1}(A_t = k, N_k(t) = i)$ does not depend on $X_{i,k}^*$ by the definition of A_t and $N_k(t)$.

Now, let $\mathcal{D}_\infty^{*'}$ be an other the tabular representation which is identical to $\mathcal{D}_\infty^*$ except the (i,k)-th entry of X_∞^* in $\mathcal{D}_\infty^*$ being replaced with an independent copy $X_{i,k}^{*'}$ from the same distribution P_k.

Since the function h is a decreasing function of $X_{i,k}^*$ while keeping all other entries in $\mathcal{D}_\infty^*$ fixed, we have that

$$\left[h(\mathcal{D}_\infty^*) - h(\mathcal{D}_\infty^{*'})\right]\left[\mathbb{1}(X_{i,k}^* \le y) - \mathbb{1}(X_{i,k}^{*'} \le y)\right] \ge 0, \tag{B.4}$$

which implies that

$$\begin{aligned} &h(\mathcal{D}_\infty^*)\left[\mathbb{1}(X_{i,k}^* \le y) - F_k(y)\right] + h(\mathcal{D}_\infty^{*'})\left[\mathbb{1}(X_{i,k}^{*'} \le y) - F_k(y)\right] + \\ &\ge h(\mathcal{D}_\infty^{*'})\left[\mathbb{1}(X_{i,k}^* \le y) - F_k(y)\right] + h(\mathcal{D}_\infty^*)\left[\mathbb{1}(X_{i,k}^{*'} \le y) - F_k(y)\right]. \end{aligned} \tag{B.5}$$

By multiplying $\mathbb{1}\left(A_t = k, N_k(t) = i\right)$ and taking expectations on both sides, we can show the inequality (B.1) hold as follows:

$$2\mathbb{E}\left[\mathbb{1}\left(A_t = k, N_k(t) = i\right) h(\mathcal{D}_\infty^*)\left[\mathbb{1}(X_{i,k}^* \le y) - F_k(y)\right]\right] \tag{B.6}$$

$$\ge 2\mathbb{E}\left[\mathbb{1}\left(A_t = k, N_k(t) = i\right) h(\mathcal{D}_\infty^*)\left[\mathbb{1}(X_{i,k}^{*'} \le y) - F_k(y)\right]\right] \tag{B.7}$$

$$= 2\mathbb{E}\left[\mathbb{1}\left(A_t = k, N_k(t) = i\right) h(\mathcal{D}_\infty^*)\right]\mathbb{E}\left[\mathbb{1}(X_{i,k}^{*'} \le y) - F_k(y)\right] \tag{B.8}$$

$$\ge 0, \tag{B.9}$$

where the first equality comes from the independence between $\mathbb{1}\left(A_t = k, N_k(t) = i\right) h(\mathcal{D}_\infty^*)$ and $X_{i,k}^{*'}$, and the second inequality holds since $\mathbb{E}\left[\mathbb{1}(X_{i,k}^{*'} \le y)\right] = F_k(y)$. □

Based on the inequality (B.1), we are ready to prove Theorem 3.1.

Proof of Theorem 3.1. First, suppose the function $\mathcal{D}_\infty^* \mapsto \mathbb{1}\left(C\right)/N_k(\mathcal{T})$ is an decreasing function of $X_{i,k}^*$ while keeping all other entries in $\mathcal{D}_\infty^*$ fixed for each i. Let $\{L_t\}_{t\in\mathbb{N}}$ be a sequence of random variables defined as follows:

$$L_t := \sum_{s=1}^{t} \frac{\mathbb{1}(C)}{N_k(\mathcal{T})}\mathbb{1}(N_k(\mathcal{T}) < \infty)\mathbb{1}\left(A_s = k\right)\left[\mathbb{1}(Y_s \le y) - F_k(y)\right], \quad \forall t \in \mathbb{N}. \tag{B.10}$$

From the inequality (B.1), we have

$$\mathbb{E}\left[L_t\right] = \sum_{s=1}^{t}\mathbb{E}\left[\frac{\mathbb{1}(C)}{N_k(\mathcal{T})}\mathbb{1}\left(A_s = k\right)\mathbb{1}(N_k(\mathcal{T}) < \infty)\left[\mathbb{1}(Y_s \le y) - F_k(y)\right]\right] \ge 0, \quad \forall t \in \mathbb{N}. \tag{B.11}$$

Note that $N_k(\mathcal{T}) := \sum_{t=1}^{\mathcal{T}} \mathbb{1}(A_t = k) = \sum_{t=1}^{\infty} \mathbb{1}(A_t = k)$ since it is understood that for $t > \mathcal{T}$, $\mathbb{1}(A_t = k) = 0$. Therefore, we know that, for each $y \in \mathbb{R}$, the sequence of random variables $\{L_t\}_{t\in\mathbb{N}}$ converges to $\left[\widehat{F}_{k,\mathcal{T}}(y) - F_k(y)\right]\mathbb{1}(C)\mathbb{1}(N_k(\mathcal{T}) < \infty)$ almost surely. Also, it can be easily checked for each $t \in \mathbb{N}$, $|L_t|$ is upper bounded by 2. Hence, from the dominated convergence theorem and the inequality (B.11), we have

$$0 \le \lim_{t\to\infty}\mathbb{E}[L_t] = \mathbb{E}\left[\left[\widehat{F}_{k,\mathcal{T}}(y) - F_k(y)\right]\mathbb{1}(C)\mathbb{1}(N_k(\mathcal{T}) < \infty)\right] \tag{B.12}$$

$$= \mathbb{E}\left[\widehat{F}_{k,\mathcal{T}}(y)\mathbb{1}(C)\mathbb{1}(N_k(\mathcal{T}) < \infty)\right] - F_k(y)\mathbb{P}(C \cap \{N_k(\mathcal{T}) < \infty\}). \tag{B.13}$$

Since $\widehat{F}_{k,\mathcal{T}}(y)\mathbb{1}(N_k(\mathcal{T})=\infty) = F_k(y)\mathbb{1}(N_k(\mathcal{T})=\infty)$ almost surely, the last inequality also implies that

$$F_k(y)\mathbb{P}(C) \leq \mathbb{E}\left[\widehat{F}_{k,\mathcal{T}}(y)\mathbb{1}(C)\right] \tag{B.14}$$

Since we assumed $\mathbb{P}(C) > 0$, by multiplying $1/\mathbb{P}(C)$ on both sides, we have

$$F_k(y) \leq \mathbb{E}\left[\widehat{F}_{k,\mathcal{T}}(y) \mid C\right], \tag{B.15}$$

as desired. The inequality (B.15) shows that the underlying distribution of arm k stochastically dominates the empirical distribution of arm k in the conditional expectation. In this case, it is well-known that for any non-decreasing integrable function f, the following inequality holds

$$E_k f \geq \mathbb{E}\left[\widehat{E}_{k,\mathcal{T}} f \mid C\right]. \tag{B.16}$$

For the completeness of the proof, we formally prove the inequality (B.16). Since f is integrable, without loss of generality, we may assume $f \geq 0$. For any $x \in \mathbb{R}$, define $f^{-1}(x) := \inf\{y : f(y) > x\}$. Since f is non-decreasing, for any probability measure P, the following equality holds

$$P\left(\{y : f(y) > x\}\right) = P\left(\left\{y : y > f^{-1}(x)\right\}\right),$$

for all but at most countably many $x \in \mathbb{R}$ which implies that

$$\begin{aligned}
E_k f &= \int_0^\infty P_k\left(\{y : f(y) > x\}\right) \mathrm{d}x && \text{(B.17)}\\
&= \int_0^\infty 1 - F_k\left(f^{-1}(x)\right) \mathrm{d}x && \text{(B.18)}\\
&\geq \int_0^\infty 1 - \mathbb{E}\left[\widehat{F}_{k,\mathcal{T}}\left(f^{-1}(x)\right) \mid C\right] \mathrm{d}x && \text{(B.19)}\\
&= \int_0^\infty \mathbb{E}\left[\widehat{P}_{k,\mathcal{T}}\left(\{y : f(y) > x\}\right) \mid C\right] \mathrm{d}x && \text{(B.20)}\\
&= \mathbb{E}\left[\widehat{E}_{k,\mathcal{T}} f \mid C\right], && \text{(B.21)}
\end{aligned}$$

where the first and last equalities come from the Fubini's theorem with the integrability condition on f, and the first inequality comes from the inequality (B.15).

From the same argument with reversed inequalities, it can be shown that if the function $\mathcal{D}_\infty^* \mapsto \mathbb{1}(C)/N_k(\mathcal{T})$ is an increasing function of $X_{i,k}^*$ while keeping all other entries in $\mathcal{D}_\infty^*$ fixed for each i, we have

$$F_k(y) \geq \mathbb{E}\left[\widehat{F}_{k,\mathcal{T}}(y) \mid C\right], \quad \forall y \in \mathbb{R}. \tag{B.22}$$

Equivalently, for any non-decreasing integrable function f, we have

$$E_k f \leq \mathbb{E}\left[\widehat{E}_k f \mid C\right], \tag{B.23}$$

which completes the proof of Theorem 3.1. □

B.1.2 Proof of Corollary 3.3

Before proving Corollary 3.3 formally, we first provide an intuition as to why, for any reasonable and efficient algorithm for the best-arm identification problem, the sample mean and empirical CDF of an arm are negative and positive biases, respectively, conditionally on the event that the arm is not chosen as the best arm.

For any $k \in [K]$ and $i \in \mathbb{N}$, let $\mathcal{D}_\infty^*$ and $\mathcal{D}_\infty^{*'}$ be two collections of all possible arm rewards and external randomness that agree with each other except $X_{i,k}^* \geq X_{i,k}^{*'}$. Since we have a smaller reward from arm k in the second scenario $\mathcal{D}_\infty^{*'}$, if $\kappa \neq k$ under the first scenario $\mathcal{D}_\infty^*$, any reasonable algorithm also would not pick the arm k as the best arm under the more unfavorable scenario $\mathcal{D}_\infty^{*'}$. In this case, we know that $\kappa \neq k$ implies $\kappa' \neq k$. Also note that any efficient algorithm should be able to exploit the more unfavorable scenario $D_\infty^{*'}$ to easily identify arm k as a suboptimal arm and choose another arm as the best one by using less samples from arm k. Therefore, we would have $N_k(\mathcal{T}) \geq N_k'(\mathcal{T}')$. As a result, we can expect that, from any reasonable and efficient algorithm, we would have $\frac{\mathbb{1}(\kappa \neq k)}{N_k(\mathcal{T})} \leq \frac{\mathbb{1}(\kappa' \neq k)}{N_k'(\mathcal{T}')}$ which implies that for each i, the function $\mathcal{D}_\infty^* \mapsto \mathbb{1}(C)/N_k(\mathcal{T})$ is a decreasing function of $X_{i,k}^*$ while keeping all other entries in $\mathcal{D}_\infty^*$ fixed. Then, from Theorem 3.1, we have that the sample mean and empirical CDF of arm k are negatively and positively biased conditionally on the event $\kappa \neq k$, respectively. Below, we formally verify that this intuition works for the lil'UCB algorithm. The proof is based on the following two facts about the lil'UCB algorithm:

- **Fact 1.** The lil'UCB algorithm has an optimistic sampling rule. That is, for any fixed , $i, t \in \mathbb{N}$ and $k \in [K]$, the function $\mathcal{D}_\infty^* \mapsto N_k(t)$ is an increasing function of $X_{i,k}^*$ while keeping all other entries in $\mathcal{D}_\infty^*$ fixed [see Fact 3 in Shin et al., 2019a].
- **Fact 2.** Let $\mathcal{D}_\infty^*$ and $\mathcal{D}_\infty^{*'}$ be two collections of all possible arm rewards and external randomness that agree with each other except in their k-th column of stacks of rewards X_∞^* and $X_\infty^{*'}$. For $j \in [K]$, let $N_j(t)$ and $N_j'(t)$ be the numbers of draws from arm j under $\mathcal{D}_\infty^*$ and $\mathcal{D}_\infty^{*'}$ respectively. Then for each $t \in \mathbb{N}$, the following implications hold for lil'UCB algorithm [see Fact 3 and Lemma 9 in Shin et al., 2019a]:

$$N_k(t) \leq N_k'(t) \Rightarrow N_j(t) \geq N_j'(t), \quad \text{for all } j \neq k,$$
$$N_k(t) \geq N_k'(t) \Rightarrow N_j(t) \leq N_j'(t), \quad \text{for all } j \neq k,$$

  which also implies that

$$N_k(t) = N_k'(t) \Rightarrow N_j(t) = N_j'(t), \quad \text{for all } j \neq k.$$

Proof of Corollary 3.3. For any given $i \in \mathbb{N}$ and $k \in [K]$, let $\mathcal{D}_\infty^*$ and $\mathcal{D}_\infty^{*'}$ be two collections of all possible arm rewards and external randomness that agree with each other except (i,k)-th entries, $X_{i,k}^*$ and $X_{i,k}^{*'}$ of their stacks of rewards. Let $(N_k(t), N_k'(t))$ denote the numbers of draws from arm k up to time t. Let $(\mathcal{T}, \mathcal{T}')$ be the stopping times and (κ, κ') be choosing functions of the lil'UCB algorithm under $\mathcal{D}_\infty^*$ and $\mathcal{D}_\infty^{*'}$ respectively.

Suppose $X_{i,k}^* \geq X_{i,k}^{*'}$. To prove the claimed bias result, it is enough to show that the function $\mathcal{D}_\infty^* \mapsto \mathbb{1}(C)/N_k(\mathcal{T})$ is a decreasing function of $X_{i,k}^*$ while keeping all other entries in $\mathcal{D}_\infty^*$ fixed which corresponds to prove the following inequality holds:

$$\frac{\mathbb{1}(\kappa \neq k)}{N_k(\mathcal{T})} \leq \frac{\mathbb{1}(\kappa' \neq k)}{N_k'(\mathcal{T}')}. \tag{B.24}$$

Note that if $\kappa = k$ or $N_k(\mathcal{T}) = \infty$, the inequality (B.24) holds trivially. Therefore, for the rest of the proof, we assume $\kappa \neq k$ and $N_k(\mathcal{T}) < \infty$.

We will first prove the inequality $N_k(\mathcal{T}) \geq N'_k(\mathcal{T}')$ holds. From Fact 1 and the assumption $X^*_{i,k} \geq X^{*'}_{i,k}$, we have $N_k(t) \geq N'_k(t)$ for any fixed $t > 0$. Then, by Fact 2, we also have $N_j(t) \leq N'_j(t)$ for any $j \neq k$. Since $\sum_{i \neq j} N_i(t) = t - N_j(t)$ for all t, we can rewrite the lil'UCB stopping rule as stopping whenever there exists $j \in [K]$ such that the inequality $N_j(t) \geq \frac{1+\lambda t}{1+\lambda}$ holds. Therefore, from the definition of the stopping rule with the fact $N_j(t) \leq N'_j(t)$ for any $t \geq 1$ and $j \neq k$, at the stopping time $\mathcal{T}$, we have

$$\frac{1+\lambda\mathcal{T}}{1+\lambda} \leq N_j(\mathcal{T}) \leq N'_j(\mathcal{T}), \tag{B.25}$$

for some $j \neq k$ which also implies that the stopping condition is also satisfied for arm j at time $\mathcal{T}$ under $\mathcal{D}^{*'}_\infty$ which implies that the stopping time under $\mathcal{D}^{*'}_\infty$ must be at most $\mathcal{T}$. Therefore we have $\mathcal{T}' \leq \mathcal{T}$. Now, since the inequality $N_k(t) \geq N'_k(t)$ holds for any $t \geq 1$, we have $N_k(\mathcal{T}) \geq N'_k(\mathcal{T})$. Finally, since $t \mapsto N'_k(t)$ is a non-decreasing function, we can conclude $N_k(\mathcal{T}) \geq N'_k(\mathcal{T}) \geq N'_k(\mathcal{T}')$.

Since we proved $N_k(\mathcal{T}) \geq N'_k(\mathcal{T}')$, to complete the proof of Corollary 3.3, it is enough to show that $\kappa \neq k$ implies $\kappa' \neq k$. We prove this statement by the proof by contradiction. Suppose $\kappa \neq k$ but $\kappa' = k$. Then, there exists $j \neq k$ such that $\kappa = j$. By the definition of $\mathcal{T}$ and κ, we know that

$$N_j(\mathcal{T}) > N_k(\mathcal{T}). \tag{B.26}$$

Similarly, we can show that

$$N'_j(\mathcal{T}') < N'_k(\mathcal{T}'). \tag{B.27}$$

It is important to note that these inequalities are strict. Note that since we draw a singe sample at each time, if $N_j(\mathcal{T}) = N_k(\mathcal{T})$ then at the time $\mathcal{T} - 1$, either arm j or k should satisfy the stopping rule which contradicts to the definition of $\mathcal{T}$.

Recall that, in Equation (B.25), we showed that if $\kappa = j$, at stopping time $\mathcal{T}$, we have

$$\frac{1+\lambda\mathcal{T}}{1+\lambda} \leq N_j(\mathcal{T}) \leq N'_j(\mathcal{T}),$$

which implies $\mathcal{T}' \leq \mathcal{T}$. By the same argument, at the stopping time $\mathcal{T}'$ with the assumption $\kappa' = k$, we have

$$\frac{1+\lambda\mathcal{T}'}{1+\lambda} \leq N'_k(\mathcal{T}') \leq N_k(\mathcal{T}'),$$

which also implies $\mathcal{T} \leq \mathcal{T}'$. From these two inequalities on stopping times, we have $\mathcal{T}' = \mathcal{T}$. Finally, by combining inequalities between pairs of N_k, N'_k, N_j, N'_j with the observation $\mathcal{T}' = \mathcal{T}$, we have

$$N'_j(\mathcal{T}') < N'_k(\mathcal{T}') \leq N_k(\mathcal{T}') = N_k(\mathcal{T}) < N_j(\mathcal{T}) \leq N'_j(\mathcal{T}) = N'_j(\mathcal{T}')$$

where the first inequality comes from the inequality (B.27). The second inequality come from $N'_k \leq N_k$. The first equality comes from $\mathcal{T}' = \mathcal{T}$ and the third inequality comes from the inequality (B.26). The last inequality comes from $N_j \leq N'_j$ and the final equality comes from $\mathcal{T} = \mathcal{T}'$.

This is a contradiction, and, therefore, $\kappa \neq k$ implies that $\kappa' \neq k$. This proves that for each i, the function $\mathcal{D}^*_\infty \mapsto \mathbb{1}(C)/N_k(\mathcal{T})$ is a decreasing function of $X^*_{i,k}$ while keeping all other entries in $\mathcal{D}^*_\infty$ fixed and, from Theorem 3.1, we can conclude that the sample mean and empirical CDF of arm k from the lil'UCB algorithm are negatively and positive biased conditionally on the event the arm k is not chosen as the best arm, respectively. □

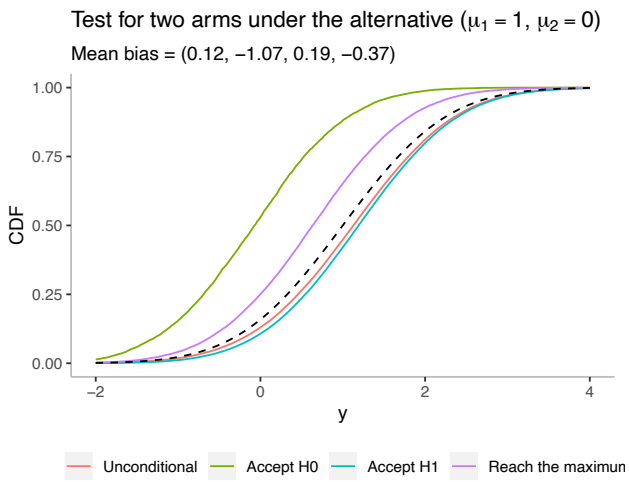

Figure B.1: *Average of conditional empirical CDFs of arm 1 from repeated sequential tests for two arms under the alternative hypothesis ($\mu_1 = 1, \mu_2 = 0$). See Section 3.2.2 for the detailed explanation about the sequential test.*

B.2 Additional Simulations results

In this section, we present additional simulation results for Section 3.2 and 3.3 which are omitted from the main part for the simple presentation.

B.2.1 Conditional bias under alternative hypothesis in Section 3.2.2

As we conducted in Section 3.2.2, we have two standard normal arms with means μ_1 and μ_2. Then, we use the following upper and lower stopping boundaries to test whether $\mu_1 \leq \mu_2$ or not:

$$U(t) := z_{\alpha/2}\sqrt{\frac{2}{t}}, \quad \text{and} \quad L(t) = -U(t), \tag{B.28}$$

where α is set to 0.2 to show the bias better. In contrast to the experiment in Section 3.2.2 in which the true means are equal to each other, in this experiment, we set $\mu_1 = 1$ and $\mu_2 = 0$ to make the alternative hypothesis is true.

Figure B.1 show the conditional and unconditional biases of the empirical CDFs and sample means for arm 1 based on 10^5 repetitions of the experiment. The dashed line corresponds to the true underlying CDF. The red line refers to the average of the empirical CDFs, and the purple line corresponds to the average of the empirical CDFs conditionally on reaching the maximal time. For these two cases, although the unconditional CDF is negatively and the conditional CDF is positively biased, these are not general phenomena and the sign of bias can be changed as we change mean parameters.

However, for the cases corresponding to accepting H_1 (blue line) and accepting H_0 (green line), we can check that signs of biases of CDFs and sample means are consistent with what Theorem 3.1 and corresponding inequalities (3.12) to (3.15) described. Also note that the bias results do not depend on whether the arms are under the null or alternative hypotheses.

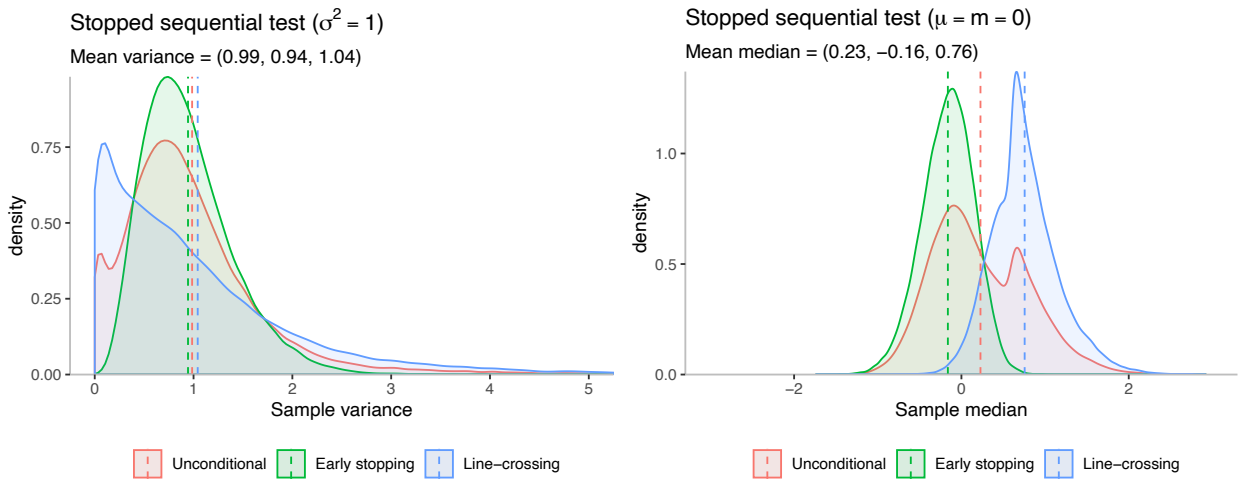

Figure B.2: *Left: Densities of observed sample variances from repeated stopped sequential test as described in Section 3.2.1. Right: Densities of observed sample median from the same repeated stopped sequential test. For both figures, vertical dashed lines correspond to averages of sample variances and medians on each conditions.*

B.2.2 Experiments on conditional biases of sample variance and median in MABs

As stated in Section 3.3, characterizing the bias of other important functionals such as sample variance and sample quantiles is an important open problem. In this subsection, we present a simulation study on the bias of sample variance and median.

For a given $n \geq 2$ i.i.d. samples $X_1, \ldots, X_n$ from a distribution P, the sample variance $\widehat{\sigma}^2$ and median $\widehat{m}$ are defined by

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \tag{B.29}$$

$$\widehat{m} = \begin{cases} \frac{1}{2}\left(X_{(n/2)} + X_{(n/2+1)}\right) & \text{if } n \text{ is even,} \\ X_{(\frac{n+1}{2})} & \text{if } n \text{ is odd,} \end{cases} \tag{B.30}$$

where $\bar{X}_n$ corresponds to the sample mean and $X_{(i)}$ refers to the i-th smallest sample.

It is well-known that for any distribution P with finite variance σ^2, the sample variance $\widehat{\sigma}^2$ is an unbiased estimator of σ^2. Also, though the sample median is not necessarily unbiased, for any symmetric distribution P including the normal distribution as a special case, the sample median is unbiased. However, for adaptively collected data from a MAB experiment, it is unclear whether the sample variance and median are unbiased or not. Furthermore, it is an open question how to characterize the bias of sample variance and median estimators if they are biased estimators.

As an initial step, we conduct the repeated sequential experiments described in Section 3.2.1 and empirically investigate the biases of the sample variance and median estimators. Figure B.2 describes a simulation study on the bias of the sample variance in the sequential testing setting of Section 3.2.1. Recall that, in this experiment, we have a stream of samples from a standard normal distribution. Each test terminates once either the number of samples reaches a fixed early stopping time $M = 10$ or the sample mean crosses the upper boundary $t \mapsto \frac{z_\alpha}{\sqrt{t}}$ with $\alpha = 0.2$.

Figure B.2 shows the conditional and unconditional distributions of the sample variance and median from 10^5 stopped sequential tests. Vertical lines correspond to averages of the sample variances and medians over repetitions of the experiment and under different conditions. For the sample variance, the simulation shows that the sample variance is negatively biased unconditionally and conditionally on the early stopping event. On the other hand, conditionally on the line-crossing event, the sample variance has a heavy right tail and is positively biased. For the sample median, we can check that, unconditionally and conditionally on the line-crossing event, the sample median is positively biased. In contrast, the sample median is negatively biased conditionally on the early stopping event. Note that, for the sample median, sizes of biases of the sample median are similar to ones from the sample means which were equal to $(0.22, -0.16, 0.75)$.

Appendix C

Appendix for Chapter 4

C.1 Examples of the Bregman divergences as a loss function

In this section, we present examples of Bregman divergences under commonly used assumptions on the underlying distribution.

Using the same notation as in Section 4.4.1, the convex conjugate of the function $\lambda \in \Lambda \mapsto \psi_\mu(\lambda) := \lambda\mu + \psi(\lambda)$ is the function ψ_μ^* on $\Lambda^* := \{x \in \mathbb{R} : \sup_{\lambda\in\Lambda} \lambda x - \psi_\mu(\lambda) < \infty\}$ given by

$$\psi_\mu^*(z) := \sup_{\lambda\in\Lambda} \lambda z - \psi_\mu(\lambda), \quad z \in \Lambda^*. \tag{C.1}$$

The Bregman divergence with respect to ψ_μ^* is then defined as

$$D_{\psi_\mu^*}(\widehat{\mu}, \mu) = \psi_\mu^*(\widehat{\mu}) - \psi_\mu^*(\mu) - \psi_\mu^{*\prime}(\mu)\,(\widehat{\mu} - \mu)\,, \quad \widehat{\mu}, \mu \in \Lambda^*. \tag{C.2}$$

Below we provide some examples demonstrating that $D_{\psi_\mu^*}(\widehat{\mu}, \mu)$ is a natural loss for the mean estimation problem when the underlying distribution is sub-ψ.

Example C.1. *If the data are generated from a sub-Gaussian distribution with parameter σ, then $\psi_\mu(\lambda)$ is defined for all $\lambda \in \mathbb{R}$ as $\psi_\mu(\lambda) := \mu\lambda + \frac{\sigma^2}{2}\lambda^2$, the Bregman divergence is defined over $\mathbb{R}$ and is equal to the scaled ℓ_2 loss:*

$$D_{\psi_\mu^*}(\widehat{\mu}, \mu) := \frac{(\widehat{\mu} - \mu)^2}{2\sigma^2}. \tag{C.3}$$

Example C.2. *If the data are generated from sub-exponential distributions with parameter (ν, α), then $\psi_\mu(\lambda)$ is defined for $\lambda \in (-1/\alpha, 1/\alpha)$ as $\psi_\mu(\lambda) = \mu\lambda + \frac{\nu^2}{2}\lambda^2$, and the Bregman divergence is defined over $\mathbb{R}$ and is given as:*

$$D_{\psi_\mu^*}(\widehat{\mu}, \mu) = \begin{cases} \frac{1}{2\nu^2}\,(\widehat{\mu} - \mu)^2\,, & \text{if } \ |\widehat{\mu} - \mu| \le \frac{\nu^2}{\alpha}, \\ \frac{1}{2\alpha}\,|\widehat{\mu} - \mu|\,, & \text{if } \ |\widehat{\mu} - \mu| > \frac{\nu^2}{\alpha}. \end{cases} \tag{C.4}$$

Example C.3. *If the data-generating distribution P satisfies the Bernstein condition*

$$\left|\mathbb{E}_{X\sim P}\,(X - \mu)^k\right| \le \frac{1}{2}k!\sigma^2 b^{k-2}, \quad \text{for } \ k = 3, 4, \ldots,$$

for some $b > 0$, where $\sigma^2 = \mathbb{E}_{X\sim P}\,(X - \mu)^2$, then, it can be shown that P is sub-ψ, where $\psi_\mu(\lambda)$ is defined for $\lambda \in (-1/b, 1/b)$ as $\psi_\mu(\lambda) = \mu\lambda + \frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}$. In this case, the Bregman divergence is defined on $\mathbb{R}$ and can be lower bounded by

$$D_{\psi_\mu^*}(\widehat{\mu}, \mu) \ge \frac{1}{2}\frac{(\widehat{\mu} - \mu)^2}{\sigma^2 + b\,|\widehat{\mu} - \mu|}. \tag{C.5}$$

Example C.4. *If the data are generated from a Bernoulli distribution, then recalling that the uncentered CGF is given by $\psi_\mu(\lambda) = \log(1 - \mu + \mu e^\lambda)$, for $\mu \in (0,1)$, the Bregman divergence is defined on $(0,1)$ and is given by*

$$D_{\psi_\mu^*}(\widehat{\mu}, \mu) = \widehat{\mu} \log \frac{\widehat{\mu}}{\mu} + (1 - \widehat{\mu}) \log \frac{1 - \widehat{\mu}}{1 - \mu}. \tag{C.6}$$

C.2 Proof of Theorem 4.6 and related statements

Recall that we assume that there exists a time t_0 such that, almost surely, $\mathcal{T} \geq \tau \geq t_0$ and $N_k(t_0) \geq 3$ for all $k \in [K]$. For ease of readability, we drop the subscript k throughout this section. We begin with the proof of the adaptive deviation inequality (4.16) of Lemma 4.7, which is a fundamental component of the proof of Theorem 4.6 and related statements.

C.2.1 Proof of Lemma 4.7

The proof strategy involves splitting the deviation event into simpler sub-events and then find exponential bounds for the probability of each sub-event. In detail, for each $t \geq 0$ and $j \geq 2$, define the events

$$\begin{aligned} F_t &:= \left\{ N(t) > e, \frac{N(t)}{4\sigma^2} \left(\widehat{\mu}_t - \mu\right)^2 > e\delta \log N(t) \right\}, \\ G_t &:= \left\{ \widehat{\mu}(t) \geq \mu \right\}, \text{ and} \\ H_t^j &:= \left\{ e^{j-1} \leq N(t) < e^j \right\}. \end{aligned}$$

We remark that the use of constant e above is purely for mathematical convenience; any other constant would have also sufficed. To bound the probability of the aforementioned events, we prove the following lemma.

Lemma C.5. *For any fixed $\delta > 0$, there exists a deterministic $\lambda_j \geq 0$ such that*

$$\left\{ F_t \cap G_t \cap H_t^j \right\} \subset \left\{ \lambda_j \left[S(t) - \mu N(t) \right] - \lambda_j^2 \sigma^2 N(t) \geq \delta j \right\}, \tag{C.7}$$

and a deterministic $\lambda_j' < 0$ such that

$$\left\{ F_t \cap G_t^c \cap H_t^j \right\} \subset \left\{ \lambda_j' \left[S(t) - \mu N(t) \right] - {\lambda_j'}^2 \sigma^2 N(t) \geq \delta j \right\}. \tag{C.8}$$

Proof of Lemma C.5. The proof borrows arguments from the proof of Theorem 11 in Garivier and Cappé [2011]. On the event $F_t \cap G_t \cap H_t^j$, since $e \leq e^{j-1} \leq N(t) < e^j$ and $v \mapsto \frac{\log v}{v}$ is non-increasing on $[e, \infty)$, we have that

$$\frac{1}{4\sigma^2} \left(\widehat{\mu}(t) - \mu\right)^2 > \frac{e\delta \log N(t)}{N(t)} \geq \frac{\delta j}{e^{j-1}} > 0. \tag{C.9}$$

Now, pick a deterministic real number $z_j \geq 0$ such that

$$\frac{1}{4\sigma^2} z_j^2 = \frac{\delta j}{e^{j-1}}.$$

Since $\widehat{\mu}(t) - \mu \geq 0$ and $x \mapsto x^2$ is an increasing function on $[0, \infty)$, our choice of z_j along with the inequalities in (C.9) implies that $\widehat{\mu}(t) - \mu \geq z_j$, on the event $F_t \cap G_t \cap H_t^j$. Define $\lambda_j := \frac{z_j}{2\sigma^2}$, then, on the event $F_t \cap G_t \cap H_t^j$, we have that

$$\lambda_j \left[\widehat{\mu}(t) - \mu\right] - \lambda_j^2 \sigma^2 \geq \lambda_j z_j - \lambda_j^2 \sigma^2 = \frac{1}{4\sigma^2} z_j^2 = \frac{\delta j}{e^{j-1}} \geq \frac{\delta j}{N(t)}.$$

Re-arranging, we get that

$$\lambda_j \left[S(t) - \mu N(t)\right] - \lambda_j^2 \sigma^2 N(t) \geq \delta j.$$

which proves the first statement in the lemma. For the second statement, set $z'_j := -z_j$. Note that, since $\widehat{\mu}(t) - \mu < 0$ and $x \mapsto x^2$ is an decreasing function on $(-\infty, 0]$, the choice of z'_j and the inequalities in (C.9) yield that $\widehat{\mu}(t) - \mu < z_j$, on the event $F_t \cap G_t^c \cap H_t^j$. Let $\lambda'_j := \frac{z'_j}{2\sigma^2}$. Then, by the same argument used for $\widehat{\mu}(t) - \mu \geq 0$ case, the second statement holds which completes the proof. □

We now return to the proof of Lemma 4.7. The adaptive deviation inequality (4.16) can be re-written as the following inequality,

$$\mathbb{P}\left(\frac{N(\tau)}{4\sigma^2}\left(\widehat{\mu}(\tau) - \mu\right)^2 \geq e\delta \log_e N(\tau)\right) \leq \frac{C_p}{\delta^p}, \tag{C.10}$$

where C_p is a constant depending only on p. To prove the above inequality, it suffices to show that it holds uniformly over time (e.g., see Lemma 3 in Howard et al. [2018b]). Therefore, below, we prove the following uniform concentration inequality:

$$\mathbb{P}\left(\exists t \in \mathbb{N} : N(t) \geq e, \frac{N(t)}{4\sigma^2}\left(\widehat{\mu}(t) - \mu\right)^2 \geq e\delta \log N(t)\right) \leq \frac{C_p}{\delta^p}. \tag{C.11}$$

The event on the left-hand side of (C.11) is equal to $\bigcup_{t=1}^{\infty} F_t$, and its probability can be bounded follows:

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{t=1}^{\infty} F_t\right) &= \mathbb{P}\left(\bigcup_{t=1}^{\infty}\bigcup_{j=2}^{\infty}\left[H_t^j \cap F_t \cap (G_t \cup G_t^c)\right]\right) \\
&= \mathbb{P}\left(\left[\bigcup_{t=1}^{\infty}\bigcup_{j=2}^{\infty}\left(H_t^j \cap F_t \cap G_t\right)\right] \cup \left[\bigcup_{t=1}^{\infty}\bigcup_{j=2}^{\infty}\left(H_t^j \cap F_t \cap G_t^c\right)\right]\right) \\
&\leq \mathbb{P}\left(\bigcup_{t=1}^{\infty}\bigcup_{j=2}^{\infty}\left(H_t^j \cap F_t \cap G_t\right)\right) + \mathbb{P}\left(\bigcup_{t=1}^{\infty}\bigcup_{j=2}^{\infty}\left(H_t^j \cap F_t \cap G_t^c\right)\right).
\end{aligned}$$

By Lemma C.5, we have that

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{t=1}^{\infty}\bigcup_{j=2}^{\infty}\left(H_t^j \cap F_t \cap G_t\right)\right) \\
&\leq \mathbb{P}\left(\exists t \geq 0, \exists j \geq 2 : \lambda_j \left[S(t) - \mu N(t)\right] - \lambda_j^2 \sigma^2 N(t) > \delta j\right) \quad \text{(by Lemma C.5)} \\
&\leq \sum_{j \geq 2} \mathbb{P}\left(\exists t \geq 0, \lambda_j \left[S(t) - \mu N(t)\right] - \lambda_j^2 \sigma^2 N(t) > \delta j\right),
\end{aligned}$$

where the last inequality stems from the union bound. To get a bound for each probability term, we use the following inequality from Khan [2009] which is a generalization of the Dubins-Savage inequality [Darling and Robbins, 1967b].

Proposition C.6 (ℓ_p-version of the Dubins-Savage inequality [Khan, 2009]). *Let $\left\{M(t) = \sum_{s=1}^{t} X_s\right\}$ be a martingale with respect to a filtration $\{\mathcal{F}_t\}_{t \geq 0}$ such that $M(0) = 0$ and $\mathbb{E}\left[X_t^{2p} \mid \mathcal{F}_{t-1}\right] < \infty$ for all*

$t \geq 1$. *Let* ν_t *be the conditional variance given by* $\nu_t = \mathbb{E}\left[X_t^2 | \mathcal{F}_{t-1}\right]$. *Then, there exist a constant* C'_p *depending only on* p *such that for any* $a \geq 0$, $b > 0$, *the following inequality holds.*

$$\mathbb{P}\left(\exists t \geq 0 : M(t) \geq a + b\sum_{s=1}^{t} \nu_s\right) \leq \frac{1}{\left(1 + ab/C'_p\right)^p}. \tag{C.12}$$

Applying inequality (C.12) with $M(t) = \lambda_j \left[S(t) - \mu N(t)\right]$, $\sum_{s=1}^{t} \nu_s = \lambda_j^2 \sigma^2 N(t)$, $a = \delta j$ and $b = 1$, we have the following bound,

$$\begin{aligned}
&\sum_{j \geq 2} \mathbb{P}\left(\exists t \geq 0, \lambda_j \left[S(t) - \mu N(t)\right] - \lambda_j^2 \sigma^2 N(t) > \delta j\right) \\
&\leq \sum_{j \geq 2} \frac{1}{\left(1 + \delta j / C'_p\right)^p} \\
&\leq \frac{C_p'^p}{\delta^p} \sum_{j \geq 2} \frac{1}{j^p} \quad := \quad \frac{C''_p}{\delta^p}.
\end{aligned}$$

Similarly, it can be shown that

$$\mathbb{P}\left(\bigcup_{t=1}^{\infty} \bigcup_{j=2}^{\infty} \left(H_t^j \cap F_t \cap G_t^c\right)\right) \leq \frac{C''_p}{\delta^p}.$$

By combining two bounds, we get that

$$\mathbb{P}\left(\exists t \geq 0 : N(t) > e, \frac{N(t)}{4\sigma^2}\left(\widehat{\mu}_t - \mu\right)^2 > e\delta \log N(t)\right) \leq \frac{2C''_p}{\delta^p},$$

which implies the desired bound on the adaptive deviation probability in (4.16) with $C_p := 2(4e)^p C''_p$. This completes the proof of Lemma 4.7.

C.2.2 Proof of Theorem 4.6

The proof of Theorem 4.6 borrows arguments from Jiao et al. [2017], like the following lower bound of D_{f_q}.

Lemma C.7. *Let* P, Q *be probability measures on* $\mathcal{X}$ *and let* $f : \mathcal{X} \mapsto \mathbb{R}$ *be a function satisfying* $\mathbb{E}_Q\left[f^p(X)\right] < \infty$ *for some* $p \geq 1$. *Then, for* q *such that* $1/p + 1/q = 1$, *we have*

$$\frac{1}{q} D_{f_q}(P||Q) \geq \mathbb{E}_P\left[f(X)\right] - \mathbb{E}_Q\left[f(X)\right] - \mathbb{E}_Q\left[\frac{|f(X)|^p}{p}\right]. \tag{C.13}$$

To apply the above inequality, we need the following bound on the expectation of the $2p$-norm of the stopped adaptive process, which is based on the adaptive deviation inequality in Lemma 4.7.

Claim C.8. *Under the assumptions of Theorem 4.6, for each* $k \in [K]$ *and for any* $\alpha \leq p$ *we have that*

$$\left\| \frac{N_k(\tau)}{\log N_k(\tau)} \left(\frac{\widehat{\mu}_k(\tau) - \mu_k}{\sigma_k}\right)^2 \right\|_\alpha \leq C_{\alpha,\epsilon}, \tag{C.14}$$

where $C_{\alpha,\epsilon}$ *is a constant depending only on* α *and* ϵ.

Proof of Claim C.8. Since $\alpha \leq p$, arm k also has a finite $2(\alpha+\epsilon)$-norm. Therefore, applying Lemma 4.7 with $p = \alpha$, we get

$$\begin{aligned}\mathbb{E}\left[\frac{N_k(\tau)}{\log N_k(\tau)}\left(\frac{\widehat{\mu}_k(\tau)-\mu_k}{\sigma_k}\right)^2\right]^{\alpha} &\leq 1+\int_1^{\infty}\mathbb{P}\left(\frac{N_k(\tau)}{\log N_k(\tau)}\left(\frac{\widehat{\mu}_k(\tau)-\mu_k}{\sigma_k}\right)^2 > \delta^{1/\alpha}\right)\mathrm{d}\delta \\ &\leq 1+C_{\alpha+\epsilon}\int_1^{\infty}\frac{1}{\delta^{1+\epsilon/\alpha}}\mathrm{d}\delta = 1+\frac{C_{\alpha+\epsilon}}{\epsilon/\alpha}.\end{aligned} \tag{C.15}$$

The claim readily follows by letting $C_{\alpha,\epsilon} := \left(1+\frac{C_{\alpha+\epsilon}}{\epsilon/\alpha}\right)^{1/\alpha}$. □

We now have all the components in place to complete the proof of Theorem 4.6. For $k \in [K]$, set $P_k = \mathcal{L}\left(\mathcal{D}_{\mathcal{T}}|\kappa=k\right)$, $Q = \mathcal{L}\left(\mathcal{D}_{\mathcal{T}}\right)$ and

$$f_k = \lambda\frac{N_k(\tau)}{\log N_k(\tau)}\left(\widehat{\mu}_k(\tau)-\mu_k\right)^2.$$

for a $\lambda > 0$. Then, from Lemma C.7, we can lower bound $I_q\left(\kappa, \mathcal{D}_{\mathcal{T}}\right)$ as follows:

$$\begin{aligned}\frac{1}{q}I_q\left(\kappa,\mathcal{D}_{\mathcal{T}}\right) &= \sum_{k=1}^{K}\mathbb{P}(\kappa=k)\left\{\frac{1}{q}D_{f_q}\left(\mathcal{L}\left(\mathcal{D}_{\mathcal{T}}|\kappa=k\right)||\mathcal{L}\left(\mathcal{D}_{\mathcal{T}}\right)\right)\right\} \\ &\geq \sum_{k=1}^{K}\mathbb{P}(\kappa=k)\left\{\mathbb{E}_{P_k}\left[f_k\right]-\mathbb{E}_Q\left[f_k\right]-\mathbb{E}_Q\left[\frac{|f_k|^p}{p}\right]\right\} \\ &= \sum_{k=1}^{K}\mathbb{P}(\kappa=k)\left\{\lambda\mathbb{E}\left[\frac{N_k(\tau)}{\log N_k(\tau)}\left(\widehat{\mu}_k(\tau)-\mu_k\right)^2 \mid \kappa=k\right]\right. \\ &\qquad \left.-\lambda\mathbb{E}\left[\frac{N_k(\tau)}{\log N_k(\tau)}\left(\widehat{\mu}_k(\tau)-\mu_k\right)^2\right]-\frac{\lambda^p}{p}\mathbb{E}\left[\frac{N_k(\tau)}{\log N_k(\tau)}\left(\widehat{\mu}_k(\tau)-\mu_k\right)^2\right]^p\right\} \\ &\geq \sum_{k=1}^{K}\mathbb{P}(\kappa=k)\left\{\lambda\mathbb{E}\left[\frac{N_k(\tau)}{\log N_k(\tau)}\left(\widehat{\mu}_k(\tau)-\mu_k\right)^2 \mid \kappa=k\right]-\left(\lambda C_{1,\epsilon}\sigma_k^2+(\lambda C_{p,\epsilon}\sigma_k^2)^p/p\right)\right\} \\ &= \lambda\mathbb{E}\left[\frac{N_\kappa(\tau)}{\log N_\kappa(\tau)}\left(\widehat{\mu}_\kappa(\tau)-\mu_\kappa\right)^2\right]-\left(\lambda C_{1,\epsilon}\|\sigma_\kappa\|_2^2+\frac{\lambda^p C_{p,\epsilon}^p}{p}\|\sigma_\kappa\|_{2p}^{2p}\right).\end{aligned}$$

Since this inequality holds for any $\lambda > 0$, we get

$$\begin{aligned}\mathbb{E}\left[\frac{N_\kappa(\tau)}{\log N_\kappa(\tau)}\left(\widehat{\mu}_\kappa(\tau)-\mu_\kappa\right)^2\right] &= C_{1,\epsilon}\|\sigma_\kappa\|_2^2+\inf_{\lambda>0}\frac{1}{\lambda}\left\{\frac{I_q\left(\kappa,\mathcal{D}_{\mathcal{T}}\right)}{q}+\frac{\lambda^p C_{p,\epsilon}^p}{p}\|\sigma_\kappa\|_{2p}^{2p}\right\} \\ &= C_{1,\epsilon}\|\sigma_\kappa\|_2^2+C_{p,\epsilon}\|\sigma_\kappa\|_{2p}^2 I_q^{1/q}\left(\kappa,\mathcal{D}_{\mathcal{T}}\right),\end{aligned}$$

thus completing the proof of the theorem.

We conclude this section with a short proof of Corollary 4.8.

C.2.3 Proof of Corollary 4.8

For any $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, Hölder's inequality along with the bound on the adaptive risk in (4.15) implies that

$$\begin{aligned}
\left[\mathbb{E}\left(\widehat{\mu}_\kappa - \mu_\kappa\right)^{2/p}\right]^p &= \left[\mathbb{E}\left(\frac{N_\kappa}{\log N_\kappa(\tau)}\right)^{-1/p}\left(\frac{N_\kappa}{\log N_\kappa(\tau)}\right)^{1/p}\left(\widehat{\mu}_\kappa - \mu_\kappa\right)^{2/p}\right]^p \\
&\leq \left[\mathbb{E}\left(\frac{N_\kappa}{\log N_\kappa(\tau)}\right)^{-q/p}\right]^{p/q} \mathbb{E}\left[\frac{N_\kappa}{\log N_\kappa(\tau)}\left(\widehat{\mu}_\kappa - \mu_\kappa\right)^2\right] \\
&\leq \left[\mathbb{E}\left(\frac{N_\kappa}{\log N_\kappa(\tau)}\right)^{-q/p}\right]^{p/q} \left[C_{1,\epsilon}\|\sigma_\kappa\|_2^2 + C_{p,\epsilon}\|\sigma_\kappa\|_{2p}^2 I_q^{1/q}\left(\kappa, \mathcal{D}_\mathcal{T}\right)\right] \\
&= \frac{1}{\widetilde{n}^{\mathrm{eff},q/p}}\left[C_{1,\epsilon}\|\sigma_\kappa\|_2^2 + C_{p,\epsilon}\|\sigma_\kappa\|_{2p}^2 I_q^{1/q}\left(\kappa, \mathcal{D}_\mathcal{T}\right)\right].
\end{aligned}$$

By setting $r := 1/p$, we infer inequality (4.18), completing the proof.

C.3 Proofs of Theorem 4.12 and related statements

The proof of Theorem 4.12 is based on the deviation inequality given in Lemma 4.13, which we prove first.

C.3.1 Proof of Lemma 4.13

The proof of the deviation inequality in Lemma 4.13 is based on the following bound on the expectation of the exponential of the stopped process. Similar versions of this bound has been exist in the literature: see, e.g., see Garivier and Cappé [2011], Howard et al. [2018a]. For the completeness, we provide the proof of the bound.

Claim C.9. *Under the assumptions of Theorem 4.12, for any $\lambda \in \Lambda$, it holds that*

$$\mathbb{E}\left[\exp\left\{\lambda\left(S_k(\mathcal{T}) - \mu_k N_k(\mathcal{T})\right) - N_k(\mathcal{T})\psi(\lambda)\right\}\right] \leq 1. \tag{C.16}$$

Proof of Claim C.9. Set $L_t^k(\lambda) := \exp\left\{\lambda\left(S_k(t) - \mu_k N_k(t)\right) - N_k(t)\psi(\lambda)\right\}$. First note that, for any $t \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\lambda\left[\left(S_k(t+1) - \mu_k N_k(t+1)\right) - \left(S_k(t) - \mu_k N_k(t)\right)\right]\right\} \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\exp\left\{\lambda \mathbb{1}(A_{t+1} = k)\left[Y_{t+1} - \mu_k\right]\right\} \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\mathbb{1}(A_{t+1} = k)\exp\left\{\lambda\left(Y_{t+1} - \mu_k\right)\right\} + \mathbb{1}(A_{t+1} \neq k) \mid \mathcal{F}_t\right] \\
&= \mathbb{1}(A_{t+1} = k)\mathbb{E}\left[\exp\left\{\lambda\left(Y_{t+1} - \mu_k\right)\right\} \mid \mathcal{F}_t\right] + \mathbb{1}(A_{t+1} \neq k) \ \text{(since } \mathbb{1}\left(A_{t+1} = k\right) \in \mathcal{F}_t.\text{)} \\
&\leq \mathbb{1}(A_{t+1} = k)\exp\left\{\psi(\lambda)\right\} + \mathbb{1}(A_{t+1} \neq k) \ \text{(since } k\text{-the distribution is sub-}\psi.\text{)} \\
&= \exp\left\{\mathbb{1}(A_{t+1} = k)\psi(\lambda)\right\} \\
&= \exp\left\{\left[N_k(t+1) - N_k(t)\right]\psi(\lambda)\right\}.
\end{aligned}$$

Thus, we obtain that

$$\begin{aligned}
&\mathbb{E}\left[L_{t+1}^k(\lambda) \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\exp\left\{\lambda\left(S_k(t+1) - \mu_k N_k(t+1)\right) - N_k(t+1)\psi(\lambda)\right\} \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\exp\left\{\lambda\left[\left(S_k(t+1) - \mu_k N_k(t+1)\right) - \left(S_k(t) - \mu_k N_k(t)\right)\right]\right\} \mid \mathcal{F}_t\right] \\
&\quad \cdot \exp\left\{\lambda\left(S_k(t) - \mu_k N_k(t)\right) - N_k(t+1)\psi(\lambda)\right\} \quad (\text{since } S_k(t), N_k(t), N_k(t+1) \in \mathcal{F}_t.) \\
&\leq \exp\left\{\left[N_k(t+1) - N_k(t)\right]\psi(\lambda)\right\} \exp\left\{\lambda\left(S_k(t) - \mu_k N_k(t)\right) - N_k(t+1)\psi(\lambda)\right\} \\
&\leq \exp\left\{\lambda\left(S_k(t) - \mu_k N_k(t)\right) - N_k(t)\psi(\lambda)\right\} \\
&= L_t^k(\lambda).
\end{aligned}$$

In particular,

$$\mathbb{E}\left[L_1^k(\lambda) \mid \mathcal{F}_0\right] = 1 := L_0^k, \ \ \forall \lambda \in \Lambda.$$

Therefore $\{L_t^k(\lambda)\}_{t\geq 0}$ is a non-negative super-martingale, and the result follows from the optional stopping theorem. □

Returning to the proof of Lemma 4.13, we first consider the case $\mathbb{P}(\mathcal{T} \leq M) = 1$ for some constant $M > 0$. Since $N_k(\mathcal{T}) \leq \mathcal{T}$, we must also have that $\mathbb{P}(N_k(\mathcal{T}) \leq M) = 1$. Next, for any $\epsilon \geq 0$ and $\lambda \in [0, \lambda_{\max}/p) \subset \Lambda$, we have

$$\begin{aligned}
\mathbb{P}\left(S_k(\mathcal{T})/N_k(\mathcal{T}) - \mu_k \geq \epsilon\right) &= \mathbb{P}\left(S_k(\mathcal{T}) \geq N_k(\mathcal{T})(\epsilon + \mu_k)\right) \\
&= \mathbb{P}\left(\exp\left\{\lambda S_k(\mathcal{T}) - \lambda(\epsilon + \mu_k)N_k(\mathcal{T})\right\} \geq 1\right) \\
&\leq \mathbb{E}\left[\exp\left\{\lambda S_k(\mathcal{T}) - \lambda(\epsilon + \mu_k)N_k(\mathcal{T})\right\}\right],
\end{aligned}$$

where in the final step we have used Markov's inequality. By using Hölder's inequality with any conjugate pairs $p, q > 1$ with $1/p + 1/q = 1$, the last term can be bounded as follows:

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\lambda\left(S_k(\mathcal{T}) - \mu_k N_k(\mathcal{T})\right) - \lambda\epsilon N_k(\mathcal{T})\right\}\right] \\
&= \mathbb{E}\left[\exp\left\{\lambda\left(S_k(\mathcal{T}) - \mu_k N_k(\mathcal{T})\right) - \frac{N_k(\mathcal{T})}{p}\psi(p\lambda)\right\} \exp\left\{N_k(\mathcal{T})\left(\frac{1}{p}\psi(p\lambda) - \lambda\epsilon\right)\right\}\right] \\
&\leq \left[\mathbb{E}\exp\left\{p\lambda\left(S_k(\mathcal{T}) - \mu_k N_k(\mathcal{T})\right) - N_k(\mathcal{T})\psi(p\lambda)\right\}\right]^{1/p} \left[\mathbb{E}\exp\left\{qN_k(\mathcal{T})\left(\frac{1}{p}\psi(p\lambda) - \lambda\epsilon\right)\right\}\right]^{1/q} \\
&\leq \left[\mathbb{E}\exp\left\{qN_k(\mathcal{T})\left(\frac{1}{p}\psi(p\lambda) - \lambda\epsilon\right)\right\}\right]^{1/q}.
\end{aligned}$$

where the last inequality follows from Claim C.9. Thus we have established the following intermediate bound on the deviation probability:

$$\mathbb{P}\left(S_k(\mathcal{T})/N_k(\mathcal{T}) - \mu_k \geq \epsilon\right) \leq \left[\mathbb{E}\exp\left\{qN_k(\mathcal{T})\left(\frac{1}{p}\psi(p\lambda) - \lambda\epsilon\right)\right\}\right]^{1/q}. \tag{C.17}$$

Since $\epsilon \geq 0$, the convex conjugate of ψ at ϵ can be written as

$$\psi^*(\epsilon) = \sup_{\lambda\in\Lambda}\left\{\lambda\epsilon - \psi(\lambda)\right\} = \sup_{\lambda\in[0,\lambda_{\max})}\left\{\lambda\epsilon - \psi(\lambda)\right\}.$$

Thus,

$$\begin{aligned}\sup_{\lambda\in[0,\lambda_{\max}/p)}\left\{\lambda\epsilon-\frac{1}{p}\psi(p\lambda)\right\}&=\frac{1}{p}\sup_{\lambda\in[0,\lambda_{\max})}\{\lambda\epsilon-\psi(\lambda)\}\\&=\frac{1}{p}\sup_{\lambda\in\Lambda}\{\lambda\epsilon-\psi(\lambda)\}\\&=\frac{1}{p}\psi^*(\epsilon).\end{aligned}$$

Using this identity, the deviation probability can be further bounded as

$$\begin{aligned}&\mathbb{P}\left(S_k(\mathcal{T})/N_k(\mathcal{T})-\mu_k\geq\epsilon\right)\\&\leq\inf_{\lambda\in[0,\lambda_{\max}/p)}\left[\mathbb{E}\exp\left\{qN_k(\mathcal{T})\left(\frac{1}{p}\psi(p\lambda)-\lambda\epsilon\right)\right\}\right]^{1/q}\\&=\left[\mathbb{E}\exp\left\{-qN_k(\mathcal{T})\sup_{\lambda\in[0,\lambda_{\max}/p)}\left(\lambda\epsilon-\frac{1}{p}\psi(p\lambda)\right)\right\}\right]^{1/q}\\&=\left[\mathbb{E}\exp\left\{-\frac{q}{p}\psi^*(\epsilon)N_k(\mathcal{T})\right\}\right]^{1/q}.\end{aligned}$$

Using the same argument, it also follows that

$$\mathbb{P}\left(S_k(\mathcal{T})/N_k(\mathcal{T})-\mu_k\leq-\epsilon\right)\leq\left[\mathbb{E}\exp\left\{-\frac{q}{p}\psi^*(-\epsilon)N_k(\mathcal{T})\right\}\right]^{1/q}.$$

Since ψ^* is a non-negative convex function with $\psi^*(0)=0$, for any $\delta\geq 0$, there exist $\epsilon_1,\epsilon_2\geq 0$ with $\psi^*(\epsilon_1)=\psi^*(-\epsilon_2)=\delta$ such that

$$\{z\in\mathbb{R}:\psi^*(z)\geq\delta\}=\{z\in\mathbb{R}:z\geq\mu_k+\epsilon_1,z\leq\mu_k-\epsilon_2\}.$$

Therefore, for any $\delta\geq 0$ and $p,q>1$ with $1/p+1/q=1$, we conclude that

$$\begin{aligned}\mathbb{P}\left(D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}),\mu_k)\geq\delta\right)&=\mathbb{P}\left(\psi^*_{\mu_k}(S_k(\mathcal{T})/N_k(\mathcal{T}))\geq\delta\right)\quad\text{(By the equality (4.25) in Fact 4.10.)}\\&\leq\mathbb{P}\left(S_k(\mathcal{T})/N_k(\mathcal{T}))-\mu_k\geq\epsilon_1\right)+\mathbb{P}\left(S_k(\mathcal{T})/N_k(\mathcal{T}))-\mu_k\leq-\epsilon_2\right)\\&\leq 2\left[\mathbb{E}\exp\left\{-\frac{q}{p}\delta N_k(\mathcal{T})\right\}\right]^{1/q}.\end{aligned}$$

For general $\mathcal{T}$, let $\mathcal{T}_M:=\min\{\mathcal{T},M\}$ for all $M>0$. Since $\mathcal{T}_M$ is a stopping time with $\mathbb{P}(\mathcal{T}_M\leq M)=1$, we have

$$\mathbb{P}\left(D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}_M),\mu_k)\geq\delta\right)\leq 2\left[\mathbb{E}\exp\left\{-\frac{q}{p}\delta N_k(\mathcal{T}_M)\right\}\right]^{1/q}\tag{C.18}$$

for any $\delta\geq 0$ and $p,q>1$ with $1/p+1/q=1$. Then, we have

$$\begin{aligned}\mathbb{P}\left(D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}),\mu_k)\geq\delta\right)&\leq\liminf_{M\to\infty}\mathbb{P}\left(D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}_M),\mu_k)\geq\delta\right)\\&\leq\liminf_{M\to\infty}2\left[\mathbb{E}\exp\left\{-\frac{q}{p}\delta N_k(\mathcal{T}_M)\right\}\right]^{1/q}\\&\leq 2\left[\mathbb{E}\exp\left\{-\frac{q}{p}\delta N_k(\mathcal{T})\right\}\right]^{1/q},\end{aligned}$$

where the first inequality comes from the Fatous's lemma and the continuity of the Bregman divergence, the second one from the inequality (C.18) and the last one from the monotone convergence theorem, along with the facts that

$$0 \leq \exp\left\{-\frac{q}{p}\delta N_k(\mathcal{T}_M)\right\} \leq 1, \ \ \forall M > 0,$$

and that $\exp\left\{-\frac{q}{p}\delta N_k(\mathcal{T}_M)\right\}$ is decreasing in M and converges almost surely to $\exp\left\{-\frac{q}{p}\delta N_k(\mathcal{T})\right\}$ as $M \to \infty$.

Finally, from the identity $q/p = q - 1$, we have that

$$\mathbb{P}\left(D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}), \mu_k)\right) \leq 2 \inf_{q>1} \left[\mathbb{E}\exp\left\{-(q-1)\delta N_k(\mathcal{T})\right\}\right]^{1/q}. \tag{C.19}$$

Since choosing $q = 1$ gives a valid, albeit trivial, bound, we can take the infimum over $q \geq 1$, which proves the first inequality in (4.30). The second inequality follows from the assumption that $N_k(\mathcal{T}) \geq b$ and the inequality

$$\begin{aligned} 2 \inf_{q\geq 1} \left[\mathbb{E}\exp\left\{-(q-1)\delta N_k(\mathcal{T})\right\}\right]^{1/q} &\leq 2 \inf_{q\geq 1} \left[\exp\left\{-(q-1)\delta b\right\}\right]^{1/q} \\ &= 2\exp\left\{-\delta b\right\}. \end{aligned}$$

This completes the proof of Lemma 4.13.

C.3.2 Proof of Theorem 4.12

In Lemma 4.13, we have established the the deviation inequality

$$\mathbb{P}\left(D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}), \mu_k) \geq \delta\right) \leq 2\left[\mathbb{E}\exp\left\{-\frac{q}{p}\delta N_k(\mathcal{T})\right\}\right]^{1/q}, \tag{C.20}$$

for any $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$. We first prove Theorem 4.12 by consider the case of $\mathbb{P}(\mathcal{T} \leq M) = 1$ for a $M > 0$. Since $N_k(\mathcal{T}) \leq \mathcal{T}$, we then have that $\mathbb{P}(N_k(\mathcal{T}) \leq M) = 1$. By using the above deviation inequality and the well-known identity $\mathbb{E}|X| = \int_0^\infty \mathbb{P}(|X| > \delta)\mathrm{d}\delta$ for any integrable random variable X, we have

$$\begin{aligned} &\mathbb{E}D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}), \mu_k) \\ &= \mathbb{E}D_{\psi^*_{\mu_k}}(S_k(\mathcal{T})/N_k(\mathcal{T}), \mu_k) \\ &= \int_0^\infty \mathbb{P}\left(D_{\psi^*_{\mu_k}}(S_k(\mathcal{T})/N_k(\mathcal{T}), \mu_k) > \delta\right)\mathrm{d}\delta \ \text{ (since the divergence is non-negative).} \\ &\leq 2\int_0^\infty \left[\mathbb{E}\exp\left\{-\frac{q}{p}\delta N_k(\mathcal{T})\right\}\right]^{1/q}\mathrm{d}\delta \ \text{ (by the deviation inequality (C.20)).} \\ &= 2\frac{ep}{b}\int_0^\infty \left[\mathbb{E}\exp\left\{-\frac{q}{p}\left(N_k(\mathcal{T}) - b/e\right)\delta\right\}\right]^{1/q}\frac{b}{ep}\exp\left\{-\frac{b}{ep}\delta\right\}\mathrm{d}\delta \\ &:= 2\frac{ep}{b}\int_0^\infty [f(\delta)]^{1/q}\, p(\delta)\mathrm{d}\delta, \end{aligned}$$

where we have set $f(\delta) = \mathbb{E}\exp\left\{-\frac{q}{p}\left(N_k(\mathcal{T}) - b/e\right)\delta\right\}$ and $p(\delta) = \frac{ep}{b}\exp\left\{-\frac{ep}{b}\delta\right\}$. Note that p is the Lebesgue density of a probability measure on $[0, \infty)$. Since $\delta \mapsto \delta^{1/q}$ is a concave function on $[0, \infty)$,

using Jensen's inequality we have that

$$\int_0^\infty [f(\delta)]^{1/q} p(\delta) \mathrm{d}\delta \leq \left[\int_0^\infty f(\delta) p(\delta) \mathrm{d}\delta \right]^{1/q}.$$

Therefore, $\mathbb{E} D_{\psi^*_{\mu_k}} (\widehat{\mu}_k(\mathcal{T}), \mu_k)$ can be further bounded by

$$\begin{aligned}
& 2\frac{ep}{b} \left[\int_0^\infty f(\delta) p(\delta) \mathrm{d}\delta \right]^{1/q} \\
&= 2\frac{ep}{b} \left[\int_0^\infty \mathbb{E} \exp\left\{ -\frac{q}{p} \left(N_k(\mathcal{T}) - b/e \right) \delta \right\} \frac{b}{ep} \exp\left\{ -\frac{b}{ep}\delta \right\} \mathrm{d}\delta \right]^{1/q} \\
&= 2\left(\frac{ep}{b}\right)^{1/p} \left[\mathbb{E} \int_0^\infty \exp\left\{ -\frac{q}{p} \left(N_k(\mathcal{T}) - \frac{b}{ep} \right) \delta \right\} \mathrm{d}\delta \right]^{1/q} \\
&= 2\left(\frac{ep}{b}\right)^{1/p} \left[\mathbb{E} \frac{1}{\frac{q}{p}\left(N_k(\mathcal{T}) - \frac{b}{ep} \right)} \right]^{1/q} \quad \left(\text{since } N_k(\mathcal{T}) \geq b > \frac{b}{ep} \right) \\
&= 2\left(\frac{e}{b}\right)^{1/p} p \left[\mathbb{E} \frac{1}{q\left(N_k(\mathcal{T}) - \frac{b}{ep} \right)} \right]^{1/q} \\
&\leq 2\left(\frac{e}{b}\right)^{1/p} p \left[\mathbb{E} \frac{1}{N_k(\mathcal{T})} \right]^{1/q},
\end{aligned}$$

where in the last inequality we have used the bound

$$N_k(\mathcal{T}) - \frac{b}{ep} = \frac{1}{p}\left(N_k(\mathcal{T}) - b/e \right) + \frac{N_k(\mathcal{T})}{q} > \frac{N_k(\mathcal{T})}{q}.$$

Thus, for any $p, q > 1$ with $1/p + 1/q = 1$, we have shown that

$$\mathbb{E} D_{\psi^*_{\mu_k}} (\widehat{\mu}_k(\mathcal{T}), \mu_k) \leq 2\left(\frac{e}{b}\right)^{1/p} p \left[\mathbb{E} \frac{1}{N_k(\mathcal{T})} \right]^{1/q}. \tag{C.21}$$

Since the above bound holds for any $p, q > 1$ with $1/p + 1/q = 1$, by taking infimum over all $p > 1$, we then have that

$$\begin{aligned}
\mathbb{E} D_{\psi^*_{\mu_k}} (\widehat{\mu}_k(\mathcal{T}), \mu_k) &\leq 2 \inf_{p>1} \left(\frac{e}{b}\right)^{1/p} p \left[\mathbb{E} \frac{1}{N_k(\mathcal{T})} \right]^{1-1/p} \\
&= \frac{2}{n^{\text{eff}}} \inf_{p>1} p \left(\frac{e n^{\text{eff}}}{b} \right)^{1/p} \\
&= \frac{2}{n^{\text{eff}}} \exp\left\{ \inf_{p>1} \left[\log p + \frac{1}{p} \log(e n^{\text{eff}}/b) \right] \right\} \\
&= \frac{2}{n^{\text{eff}}} \exp\left\{ \log\log(e n^{\text{eff}}/b) + 1 \right\} \\
&= 2e \frac{1 + \log(n_k^{\text{eff}}/b)}{n_k^{\text{eff}}},
\end{aligned}$$

where the second equality is justified by the continuity of the exponential and logarithmic functions. The third equality follows from the fact that if $a \geq e$,

$$\log p + \frac{1}{p}\log a \geq \log\log a + 1, \quad \forall p \geq 1,$$

with equality if and only if $p = \log a$. Since, by assumption $N_k(\mathcal{T}) \geq b$, we have that $n_k^{\text{eff}} \geq b$ and therefore we can set $p = \log(en^{\text{eff}}/b) \geq 1$. Thus, the first part of the claimed bound on the risk in (4.28) is proven.

To prove the second part of the upper bound, we use the deviation inequality in a slightly different way which is motivated from the proof of Theorem 12.1. in Peña et al. [2008]. Specifically, for any $\epsilon > 0$ and $r > 1$, we have

$$\begin{aligned}
&\mathbb{E}D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}), \mu_k) \\
&= \mathbb{E}D_{\psi^*_{\mu_k}}(S_k(\mathcal{T})/N_k(\mathcal{T}), \mu_k) \\
&= \int_0^\infty \mathbb{P}\left(D_{\psi^*_{\mu_k}}(S_k(\mathcal{T})/N_k(\mathcal{T}), \mu_k) > \delta\right) \mathrm{d}\delta \\
&\leq \epsilon + \int_\epsilon^\infty \mathbb{P}\left(D_{\psi^*_{\mu_k}}(S_k(\mathcal{T})/N_k(\mathcal{T}), \mu_k) > \delta\right) \mathrm{d}\delta \\
&\leq \epsilon + 2\int_\epsilon^\infty \left[\mathbb{E}\exp\left\{-\frac{q}{p}\delta N_k(\mathcal{T})\right\}\right]^{1/q} \mathrm{d}\delta \quad \text{(by the deviation inequality (C.20))}. \\
&= \epsilon + 2\int_\epsilon^\infty \delta^{-r}\left[\mathbb{E}\delta^{qr}\exp\left\{-\frac{q}{p}\delta N_k(\mathcal{T})\right\}\right]^{1/q} \mathrm{d}\delta \\
&\leq \epsilon + 2\int_\epsilon^\infty \delta^{-r}\left[\mathbb{E}\sup_{\tau>0}\tau^{qr}\exp\left\{-\frac{q}{p}\tau N_k(\mathcal{T})\right\}\right]^{1/q} \mathrm{d}\delta.
\end{aligned}$$

It can be easily checked that the supremum is achieved at $\tau = \frac{pr}{N_k(\mathcal{T})}$. Therefore we have that

$$\sup_{\tau>0}\tau^{qr}\exp\left\{-\frac{q}{p}\tau N_k(\mathcal{T})\right\} = \left(\frac{pr}{N_k(\mathcal{T})}\right)^{qr} e^{-qr},$$

which implies that

$$\begin{aligned}
\mathbb{E}D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}), \mu_k) &\leq \epsilon + 2\int_\epsilon^\infty \delta^{-r}\left[\mathbb{E}\sup_{\delta>0}\delta^{qr}\exp\left\{-\frac{q}{p}\delta N_k(\mathcal{T})\right\}\right]^{1/q} \mathrm{d}\delta \\
&= \epsilon + 2\int_\epsilon^\infty \delta^{-r}\left[\mathbb{E}\left(\frac{pr}{N}\right)^{qr} e^{-qr}\right]^{1/q} \mathrm{d}\delta \\
&= \epsilon + 2\left(\frac{pr}{e}\right)^r \left[\mathbb{E}\left(\frac{1}{N^{qr}}\right)\right]^{1/q} \int_\epsilon^\infty \delta^{-r}\mathrm{d}\delta \\
&= \epsilon + \frac{2}{(r-1)\epsilon^{r-1}}\left(\frac{pr}{en^{\text{eff},qr}}\right)^r.
\end{aligned}$$

Since the above bound holds for any $\epsilon > 0$, by taking infimum on the RHS over $\epsilon > 0$, we have the following upper bound.

$$\mathbb{E}D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(\mathcal{T}), \mu_k) \leq \frac{2^{1/r}pr^2}{e(r-1)}\frac{1}{n^{\text{eff},qr}}.$$

By setting $r' = qr$, we can write the above inequality as

$$\mathbb{E}D_{\psi^*_{\mu_k}}\left(\widehat{\mu}_k(\mathcal{T}), \mu_k\right) \leq C_{q,r'} \frac{1}{n^{\text{eff},r'}}, \tag{C.22}$$

where

$$C_{q,r'} = \frac{2^{q/r'}}{e} \frac{r'^2}{(r'-q)(q-1)}.$$

Since this upper bound holds for any choice of $r' > q > 1$, the second part of the upper bound is proved.

For general $\mathcal{T}$, let $\mathcal{T}_M := \min\{\mathcal{T}, M\}$ for all $M \geq t_0$. Since $\mathcal{T}_M$ is a stopping time with $T_M \geq t_0$ and $\mathbb{P}(\mathcal{T}_M \leq M) = 1$, we have that

$$\mathbb{E}D_{\psi^*_{\mu_k}}\left(\widehat{\mu}_k(\mathcal{T}_M), \mu_k\right) \leq \min\left\{2e\frac{1+\log(n_k^{\text{eff}\wedge M}/b)}{n_k^{\text{eff}}}, \inf_{r>1}\frac{C_r}{n_k^{\text{eff}\wedge M,r}}\right\}, \quad \forall M \geq t_0, \tag{C.23}$$

where $n_k^{\text{eff}\wedge M,r}$ is the corresponding effective sample size $\left[\mathbb{E}\left[1/N_k^r(\mathcal{T}_M)\right]\right]^{-r}$ with $n_k^{\text{eff}\wedge M} = n_k^{\text{eff}\wedge M,1}$. Then, we have

$$\begin{aligned}
\mathbb{E}D_{\psi^*_{\mu_k}}\left(\widehat{\mu}_k(\mathcal{T}), \mu_k\right) &\leq \liminf_{M\to\infty} \mathbb{E}D_{\psi^*_{\mu_k}}\left(\widehat{\mu}_k(\mathcal{T}_M), \mu_k\right) \\
&\leq \liminf_{M\to\infty} \min\left\{2e\frac{1+\log(n_k^{\text{eff}\wedge M}/b)}{n_k^{\text{eff}}}, \inf_{r>1}\frac{C_r}{n_k^{\text{eff}\wedge M,r}}\right\} \\
&\leq \min\left\{2e\frac{1+\log(n_k^{\text{eff}}/b)}{n_k^{\text{eff}}}, \inf_{r>1}\frac{C_r}{n_k^{\text{eff},r}}\right\},
\end{aligned}$$

as desired, where the first inequality comes from Fatous's lemma, the second one follows from the inequality (C.23) and the last one comes from the monotone convergence theorem along with the facts that $0 \leq 1/N_k^r(\mathcal{T}_M) \leq 1/b^r$ for all $M \geq t_0$ and that $\{1/N_k^r(\mathcal{T}_M)\}_{M\geq t_0}$ is a non-negative decreasing sequence converging to $1/N_k^r(\mathcal{T})$ almost surely which also implies $n_k^{\text{eff}\wedge M,r} \to n_k^{\text{eff},r}$ as $M \to \infty$.

The proof of Theorem 4.12 is completed. In the following subsection, we present a simple proof of Corollary 4.14.

C.3.3 Proof of Corollary 4.14

By the equation (4.25), we have that

$$D_{\psi^*_{\mu_k}}\left(\widehat{\mu}_k(\mathcal{T}), \mu_k\right) = \psi^*_{\mu_k}\left(\widehat{\mu}_k(\mathcal{T})\right) = \psi^*\left(\widehat{\mu}_k(\mathcal{T}) - \mu_k\right).$$

Since ψ^* is convex, applying the Jensen's inequality to the risk bound in the equation (4.28) of Theorem 4.12, we get that

$$\begin{aligned}
\psi^*\left(\mathbb{E}\left[\widehat{\mu}_k(\mathcal{T}) - \mu_k\right]\right) &\leq \mathbb{E}\left[\psi^*\left(\widehat{\mu}_k(\mathcal{T}) - \mu_k\right)\right] \\
&= \mathbb{E}D_{\psi^*_{\mu_k}}\left(\widehat{\mu}_k(\mathcal{T}), \mu_k\right) \leq U_{k,b}.
\end{aligned}$$

If the bias $\mathbb{E}\left[\widehat{\mu}_k(\mathcal{T}) - \mu_k\right]$ is positive, ψ^* can be replaced with ψ^*_+, which implies that

$$\psi^*_+\left(\mathbb{E}\left[\widehat{\mu}_k(\mathcal{T}) - \mu_k\right]\right) \leq U_{k,b}.$$

Since ψ_+^* is an increasing and invertible function, we get the desired upper bound on bias, namely

$$\mathbb{E}\left[\widehat{\mu}_k(\mathcal{T}) - \mu_k\right] \leq {\psi_+^*}^{-1}\left(U_{k,b}\right).$$

Applying the same argument to the case of a negative bias, we arrive at the analogous lower bound

$$-{\psi_-^*}^{-1}\left(U_{k,b}\right) \leq \mathbb{E}\left[\widehat{\mu}_k(\mathcal{T}) - \mu_k\right].$$

This completes the proof of the expression (4.32).

If ψ^* is symmetric around zero, $\psi^*(z) = \psi_+^*(|z|)$ for all $z \in \Lambda^*$. Therefore, by the same steps,

$$\begin{aligned}
\psi_+^*\left(\mathbb{E}\left|\widehat{\mu}_k(\mathcal{T}) - \mu_k\right|\right) &\leq \mathbb{E}\left[\psi_+^*\left(\left|\widehat{\mu}_k(\mathcal{T}) - \mu_k\right|\right)\right] \\
&= \mathbb{E}\left[\psi^*\left(\widehat{\mu}_k(\mathcal{T}) - \mu_k\right)\right] \\
&= \mathbb{E} D_{\psi_{\mu_k}^*}\left(\widehat{\mu}_k(\mathcal{T}), \mu_k\right) \leq U_{k,b}.
\end{aligned}$$

Applying ${\psi_+^*}^{-1}$ to the both sides, we arrive at the bound on the expected ℓ_1 loss given in (4.33) which completes the proof.

C.4 Proofs of Theorem 4.16 and related statements

In this section, Theorem 4.16 and related statements are proved. Recall that we assume that there exists a time t_0 such that, almost surely, $\mathcal{T} \geq \tau \geq t_0$ and $N_k(t_0) \geq b \geq 3$ for all $k \in [K]$. Let $h : [\log b, \infty) \to [1, \infty)$ be a non-decreasing function such that $\sum_{j=1}^{\infty} \frac{1}{h(\log b + j)} \leq 1$ and $v \in [b, \infty) \mapsto \frac{\log h(\log v)}{v}$ is non-increasing. It can be easily checked that the function $h_b(x) := x^2/\log b$ satisfies the condition above.

For ease of readability, we drop the subscript k throughout this section. We first provide the proof to the adaptive deviation inequality of Lemma 4.18, which is a fundamental component of the proof of Theorem 4.16 and related statements.

C.4.1 Proof of Lemma 4.18

The proof strategy of the adaptive deviation inequality (4.42) is based on splitting the deviation event into simpler sub-events and then find exponential bounds for the probability of each sub-event. In detail, for each $t \geq 0$ and $j \in \log b + \mathbb{N} := \{\log b + i : i \in \mathbb{N}\}$, define the events

$$\begin{aligned}
F_t &:= \left\{N(t) \geq b, N(t)\psi^*(\widehat{\mu}_t - \mu) > e\left(\delta + \log h\left(\log N(t)\right)\right)\right\}, \\
G_t &:= \left\{\widehat{\mu}(t) \geq \mu\right\}, \text{ and} \\
H_t^j &:= \left\{e^{j-1} \leq N(t) < e^j\right\}.
\end{aligned}$$

To bound the probability of these events, we rely on the following result, which we establish using arguments borrowed from the proof of Theorem 11 in Garivier and Cappé [2011], See also Garivier [2013].

Lemma C.10. *Let $h : [\log b, \infty) \to [1, \infty)$ be a non-decreasing function which makes $v \mapsto \frac{\log h(\log v)}{v}$ non-increasing on $[b, \infty)$. Then, for any fixed $\delta > 0$, there exist a deterministic $\lambda_j \geq 0$ such that*

$$\left\{F_t \cap G_t \cap H_t^j\right\} \subset \left\{\lambda_j\left[S(t) - \mu N(t)\right] - \psi(\lambda_j)N(t) \geq \delta + \log h(j)\right\}, \tag{C.24}$$

and a deterministic $\lambda_j' < 0$ such that

$$\left\{F_t \cap G_t^c \cap H_t^j\right\} \subset \left\{\lambda_j'\left[S(t) - \mu N(t)\right] - \psi(\lambda_j')N(t) \geq \delta + \log h(j)\right\}. \tag{C.25}$$

Proof of Lemma C.10. On the event $F_t \cap G_t \cap H_t^j$, since $b \leq e^{j-1} \leq N(t) < e^j$ and $v \mapsto \frac{\log h(\log v)}{v}$ is non-increasing on $[b, \infty)$, we have that

$$\psi^* \left(\widehat{\mu}(t) - \mu\right) > \frac{e\delta}{N(t)} + \frac{e \log h \left(\log N(t)\right)}{N(t)} \geq \frac{\delta}{e^{j-1}} + \frac{\log h(j)}{e^{j-1}} > 0. \tag{C.26}$$

Since, by assumption, ψ is a non-negative convex function such that $\psi(0) = \psi'(0) = 0$. its convex conjugate ψ^* is an increasing function on $[0, \infty)$ with $\psi^*(0) = 0$. Therefore, we can pick a deterministic real number $z_j \geq 0$ such that

$$\psi^*(z_j) = \frac{\delta}{e^{j-1}} + \frac{\log h(j)}{e^{j-1}}.$$

Note that, since $\widehat{\mu}(t) - \mu \geq 0$ and ψ^* is an increasing function on $[0, \infty)$, our choice of z_j along with the inequalities in (C.26) implies that $\widehat{\mu}(t) - \mu \geq z_j$, on the event $F_t \cap G_t \cap H_t^j$.

Let λ_j be the convex conjugate of z_j with respect to ψ, which is given by

$$\lambda_j = \arg\max_{\lambda \in \Lambda} \lambda z_j - \psi(\lambda) = \psi^*(z_j).$$

Since $z_j \geq 0$, λ_j is also non-negative. Therefore, on the event $F_t \cap G_t \cap H_t^j$, we have that

$$\begin{aligned} \lambda_j \left[\widehat{\mu}(t) - \mu\right] - \psi(\lambda_j) &\geq \lambda_j z_j - \psi(\lambda_j) \\ &= \psi^*(z_j) \\ &= \frac{\delta}{e^{j-1}} + \frac{\log h(j)}{e^{j-1}} \\ &\geq \frac{\delta}{N(t)} + \frac{\log h(j)}{N(t)}. \end{aligned}$$

Re-arranging, we get that

$$\lambda_j \left[S(t) - \mu N(t)\right] - \psi(\lambda_j) N(t) \geq \delta + \log h(j).$$

which proves the first statement in the lemma. For the second statement, since ψ^* is a decreasing function on $(-\infty, 0]$ with $\psi^*(0) = 0$ we can pick a deterministic real number $z_j' < 0$ such that

$$\psi^*(z_j') = \frac{\delta}{e^{j-1}} + \frac{\log h(j)}{e^{j-1}}.$$

Note that, since $\widehat{\mu}(t) - \mu < 0$ and ψ^* is an decreasing function on $(-\infty, 0]$, the choice of z_j' and the inequalities in (C.26) yield that $\widehat{\mu}(t) - \mu < z_j$, on the event $F_t \cap G_t^c \cap H_t^j$. Let λ_j' be the convex conjugate of z_j'. Since $z_j' < 0$, it is also the case that $\lambda_j' < 0$. Then, by the same argument used for $\widehat{\mu}(t) - \mu \geq 0$ case, the second statement holds which completes the proof. □

Now, we continue to prove Lemma 4.18. In Fact 4.10, we showed that $D_{\psi_\mu^*}(\widehat{\mu}, \mu) = \psi^*(\widehat{\mu} - \mu)$. Thus, the adaptive deviation inequality (4.42) can be re-stated in the general setting as the inequality

$$\mathbb{P}\left(\frac{N(\tau)}{\log h \left(\log N(\tau)\right)} \psi^* \left(\widehat{\mu}(\tau) - \mu\right) \geq C_{h,b}\delta\right) \leq 2 \exp\left\{-\delta\right\}, \quad \forall \delta \geq 1, \tag{C.27}$$

where $C_{h,b} := e\left(1 + \frac{1}{\log h(\log b)}\right)$, which can be derived from the following more general inequality.

$$\begin{aligned} &\mathbb{P}\left(N(\tau)\psi^* \left(\widehat{\mu}(\tau) - \mu\right) \geq e\left(\delta + \log h \left(\log N(\tau)\right)\right)\right) \\ &\leq 2 \exp\left\{-\delta\right\}, \quad \forall \delta \geq 0. \end{aligned} \tag{C.28}$$

To show the above bound, it is sufficient to show that the inequality holds uniformly for all time. (e.g., see Lemma 3 in Howard et al. [2018b]). Therefore, in this proof, we prove the following uniform concentration inequality:

$$\begin{aligned}&\mathbb{P}\left(\exists t \in \mathbb{N} : N(t) \geq b, N(t)\psi^*\left(\widehat{\mu}(t) - \mu\right) \geq e\left(\delta + \log h\left(\log N(t)\right)\right)\right) \\ &\leq 2\exp\left\{-\delta\right\}, \quad \forall \delta \geq 0.\end{aligned} \tag{C.29}$$

The event on the left-hand side of (C.29) is equal to $\bigcup_{t=1}^{\infty} F_t$, and its probability can be bounded as follows:

$$\begin{aligned}\mathbb{P}\left(\bigcup_{t=1}^{\infty} F_t\right) &= \mathbb{P}\left(\bigcup_{t=1}^{\infty} \bigcup_{j \in \log b + \mathbb{N}} \left[H_t^j \cap F_t \cap \left(G_t \cup G_t^c\right)\right]\right) \\ &= \mathbb{P}\left(\left[\bigcup_{t=1}^{\infty} \bigcup_{j \in \log b + \mathbb{N}} \left(H_t^j \cap F_t \cap G_t\right)\right] \cup \left[\bigcup_{t=1}^{\infty} \bigcup_{j \in \log b + \mathbb{N}} \left(H_t^j \cap F_t \cap G_t^c\right)\right]\right) \\ &\leq \mathbb{P}\left(\bigcup_{t=1}^{\infty} \bigcup_{j \in \log b + \mathbb{N}} \left(H_t^j \cap F_t \cap G_t\right)\right) + \mathbb{P}\left(\bigcup_{t=1}^{\infty} \bigcup_{j \in \log b + \mathbb{N}} \left(H_t^j \cap F_t \cap G_t^c\right)\right).\end{aligned}$$

For each λ and t, define $L_t(\lambda) := \exp\left\{\lambda\left(S(t) - \mu N(t)\right) - N(t)\psi(\lambda)\right\}$. We show in the proof of Lemma 4.13, that, for each $\lambda \in \Lambda$, $\{L_t(\lambda)\}_{t \geq 0}$ is a non-negative super-martingale with $\mathbb{E}L_0(\lambda) = 1$. By Lemma C.10, we have that

$$\begin{aligned}&\mathbb{P}\left(\bigcup_{t=1}^{\infty} \bigcup_{j=\log b + \mathbb{N}}^{\infty} \left(H_t^j \cap F_t \cap G_t\right)\right) \\ &\leq \mathbb{P}\left(\exists t \geq 0, \exists j \in \log b + \mathbb{N} : \lambda_j\left[S(t) - \mu N(t)\right] - \psi(\lambda_j)N(t) > \delta + \log h(j)\right) \quad \text{(by Lemma C.10.)} \\ &= \mathbb{P}\left(\exists t \geq 0, \exists j \in \log b + \mathbb{N} : \exp\left(\lambda_j\left[S(t) - \mu N(t)\right] - \psi(\lambda_j)N(t)\right) > h(j)e^{\delta}\right) \\ &= \mathbb{P}\left(\exists t \geq 0, \exists j \in \log b + \mathbb{N} : L_t(\lambda_j) > h(j)e^{\delta}\right) \\ &\leq \sum_{j \in \log b + \mathbb{N}} \mathbb{P}\left(\sup_{t \geq 0} L_t(\lambda_j) > h(j)e^{\delta}\right),\end{aligned}$$

where the last inequality stems from the union bound. Since $\{L_t\}_{t \geq 0}$ is a non-negative super-martingale with $L_0 = 1$, by applying Ville's maximal inequality [Ville, 1939], we conclude that

$$\sum_{j \in \log b + \mathbb{N}} \mathbb{P}\left(\sup_{t \geq 0} L_t(\lambda_j) > h(j)e^{\delta}\right) \leq e^{-\delta} \sum_{j=1}^{\infty} \frac{1}{h(\log b + j)} \leq e^{-\delta}.$$

Similarly, it can be shown

$$\mathbb{P}\left(\bigcup_{t=1}^{\infty} \bigcup_{j \in \log b + \mathbb{N}} \left(H_t^j \cap F_t \cap G_t^c\right)\right) \leq e^{-\delta}.$$

By combining two bounds, we get that

$$\mathbb{P}\left(\exists t \geq 0 : N(t) \geq b, N(t)\psi^*(\widehat{\mu}_t - \mu) > e\left(\delta + \log h\left(\log N(t)\right)\right)\right) \leq 2e^{-\delta},$$

which implies the desired bound on the adaptive deviation probability in (C.27).

C.4.2 Proof of Theorem 4.16

The proof of Theorem 4.16 relies on the Donsker-Varadhan representation of the KL divergence and arguments in Russo and Zou [2016], Jiao et al. [2017]. For completeness, we cite the following form of Donsker-Varadhan representation theorem [see, e.g. Donsker and Varadhan, 1983, Jiao et al., 2017]:

Lemma C.11. *Let P, Q be probability measures on $\mathcal{X}$ and let $\mathcal{C}$ denote the set of functions $f : \mathcal{X} \mapsto \mathbb{R}$ such that $\mathbb{E}_Q\left[e^{f(X)}\right] < \infty$. If $D_{KL}(P||Q) < \infty$ then for every $f \in \mathcal{C}$ the expectation $\mathbb{E}_P\left[f(X)\right]$ exists and furthermore*

$$D_{KL}(P||Q) = \sup_{f \in \mathcal{C}} \mathbb{E}_P\left[f(X)\right] - \log \mathbb{E}_Q\left[e^{f(X)}\right], \tag{C.30}$$

where the supremum is attained when $f = \log \frac{\mathrm{d}P}{\mathrm{d}Q}$.

To apply Donsker-Varadhan representation, we need the following bound on the expectation of the exponentiated stopped adaptive process, which is based on the adaptive deviation inequality in Lemma 4.18.

Claim C.12. *Under the assumptions of Theorem 4.12, for each $k \in [K]$ we have that*

$$\mathbb{E} \exp\left\{\frac{1}{2C_{h,b}}\left[\frac{N_k(\tau)}{\log h\left(\log N_k(\tau)\right)} D_{\psi^*_{\mu_k}}(\widehat{\mu}_k, \mu_k)\right]\right\} \leq 3.46. \tag{C.31}$$

Proof of Claim C.12. Using the inequality

$$\begin{aligned}
\mathbb{E}e^X = \mathbb{E}\int_{-\infty}^{\infty} \mathbb{1}(\delta < X) e^{\delta} \mathrm{d}\delta &= \int_{-\infty}^{\infty} \mathbb{P}(X > \delta) e^{\delta} \mathrm{d}\delta \\
&\leq \int_{1}^{\infty} \mathbb{P}(X > \delta) e^{\delta} \mathrm{d}\delta + \int_{-\infty}^{1} e^{\delta} \mathrm{d}\delta \\
&= \int_{1}^{\infty} \mathbb{P}(X > \delta) e^{\delta} \mathrm{d}\delta + e,
\end{aligned} \tag{C.32}$$

we can bound the left hand side of (C.31) as follows:

$$\begin{aligned}
&\mathbb{E} \exp\left\{\frac{1}{2C_{h,b}}\left[\frac{N_k(\tau)}{\log h\left(\log N_k(\tau)\right)} D_{\psi^*_{\mu_k}}(\widehat{\mu}_k, \mu_k)\right]\right\} \\
&\leq e + \int_{1}^{\infty} \mathbb{P}\left(\frac{1}{2C_{h,b}}\left[\frac{N_k(\tau)}{\log h\left(\log N_k(\tau)\right)} D_{\psi^*_{\mu_k}}(\widehat{\mu}_k, \mu_k)\right] > \delta\right) e^{\delta} \mathrm{d}\delta \\
&= e + \int_{1}^{\infty} \mathbb{P}\left(\frac{N_k(\tau)}{\log h\left(\log N_k(\tau)\right)} D_{\psi^*_{\mu_k}}(\widehat{\mu}_k, \mu_k) > 2C_{h,b}\delta\right) e^{\delta} \mathrm{d}\delta.
\end{aligned}$$

By applying the adaptive deviation inequality (4.42) Lemma 4.18, the last term can be further bounded by $e + \int_1^{\infty} 2e^{-\delta} \mathrm{d}\delta = e + 2/e < 3.46$. The claimed result readily follows. □

Coming back to the proof of Theorem 4.16, for each $k \in [K]$, set $P_k = \mathcal{L}\left(\mathcal{D}_{\mathcal{T}} | \kappa = k\right)$, $Q = \mathcal{L}\left(\mathcal{D}_{\mathcal{T}}\right)$ and

$$f_k = \frac{1}{2C_{h,b}}\left[\frac{N_k(\tau)}{\log h\left(\log N_k(\tau)\right)} D_{\psi^*_{\mu_k}}(\widehat{\mu}_k, \mu_k)\right].$$

Then, from the Donsker-Varadhan representation, we can lower bound the mutual information between the

adaptive query κ and the data $\mathcal{D}_{\mathcal{T}}$ in the following way:

$$
\begin{aligned}
I(\kappa; \mathcal{D}_{\mathcal{T}}) &= \sum_{k=1}^{K} \mathbb{P}(\kappa = k) D_{KL}\left(\mathcal{L}\left(\mathcal{D}_{\mathcal{T}} | \kappa = k\right) || \mathcal{L}\left(\mathcal{D}_{\mathcal{T}}\right)\right) \\
&\geq \sum_{k=1}^{K} \mathbb{P}(\kappa = k) \mathbb{E}_{P_k}\left[f_k\right] - \log \mathbb{E}_{Q}\left[e^{f_k}\right] \\
&= \sum_{k=1}^{K} \mathbb{P}(\kappa = k) \left\{ \frac{1}{2C_{h,b}} \mathbb{E}\left[\frac{N_k(\tau)}{\log h\left(\log N_k(\tau)\right)} D_{\psi^*_{\mu_k}}(\widehat{\mu}_k, \mu_k) \mid \kappa = k \right] \right. \\
&\qquad \left. - \log \mathbb{E}\left[\exp\left\{ \frac{1}{2C_{h,b}} \left[\frac{N_k(\tau)}{\log h\left(\log N_k(\tau)\right)} D_{\psi^*_{\mu_k}}(\widehat{\mu}_k, \mu_k) \right] \right\} \right] \right\} \\
&\geq \sum_{k=1}^{K} \mathbb{P}(\kappa = k) \left\{ \frac{1}{2C_{h,b}} \mathbb{E}\left[\frac{N_k(\tau)}{\log h\left(\log N_k(\tau)\right)} D_{\psi^*_{\mu_k}}(\widehat{\mu}_k, \mu_k) \mid \kappa = k \right] - \log 3.46 \right\} \\
&\geq \frac{1}{2C_{h,b}} \mathbb{E}\left[\frac{N_\kappa(\tau)}{\log h\left(\log N_\kappa(\tau)\right)} D_{\psi^*_{\mu_\kappa}}(\widehat{\mu}_\kappa, \mu_\kappa) \right] - 1.25,
\end{aligned}
$$

where the second inequality is due to the inequality (C.31) in Claim C.12. The risk bound (4.39) now follows from rearranging with $h(x) = x^2 / \log b$ and $C_b := 4C_{h,b}$ which completes the proof.

The only remaining proof in this section is that of Corollary 4.19, which we present below.

C.4.3 Proof of Corollary 4.19

For any $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, Hölder's inequality along with the bound on the adaptive risk in (4.39) implies that

$$
\begin{aligned}
\left[\mathbb{E} D^{1/p}_{\psi^*_{\mu_\kappa}}(\widehat{\mu}_\kappa, \mu_\kappa)\right]^p &= \left[\mathbb{E}\left(\frac{N_\kappa(\tau)}{\log\log N_\kappa(\tau)} \right)^{-1/p} \left(\frac{N_\kappa(\tau)}{\log\log N_\kappa(\tau)} \right)^{1/p} D^{1/p}_{\psi^*_{\mu_\kappa}}(\widehat{\mu}_\kappa, \mu_\kappa) \right]^p \\
&\leq \left[\mathbb{E}\left(\frac{N_\kappa(\tau)}{\log\log N_\kappa(\tau)} \right)^{-q/p} \right]^{p/q} \mathbb{E}\left[\frac{N_\kappa(\tau)}{\log\log N_\kappa(\tau)} D_{\psi^*_{\mu_\kappa}}(\widehat{\mu}_\kappa, \mu_\kappa) \right] \\
&\leq \left[\mathbb{E}\left(\frac{N_\kappa(\tau)}{\log\log N_\kappa(\tau)} \right)^{-q/p} \right]^{p/q} C_b \left[I(\kappa; \mathcal{D}_{\mathcal{T}}) + 1.25 \right] \\
&= \frac{C_b}{\widetilde{\widetilde{n}}^{\mathrm{eff}, q/p}} \left[I(\kappa; \mathcal{D}_{\mathcal{T}}) + 1.25 \right].
\end{aligned}
$$

By setting $r := 1/p$, we proves the inequality (4.44), as desired.

C.5 Proofs of propositions and facts

In this section, we provide formal proofs of propositions and facts which are omitted in the main text.

C.5.1 Proof of Proposition 4.2

For any $k \in [K]$, the strong law of large numbers and Theorem 2.1. in Gut [2009] implies that

$$
\text{if } \; N_k(\tau_t) \overset{a.s.}{\to} \infty \; \text{ as } \; t \to \infty \; \text{ then } \; \widehat{\mu}_k(\tau_t) \overset{a.s.}{\to} \mu_k \; \text{ as } \; t \to \infty. \tag{C.33}
$$

For each $k \in [K]$, define events E_k and F_k such that

$$E_k = (\widehat{\mu}_k(\tau_t) \to \mu_k \ \text{ as } \ t \to \infty)\,, \tag{C.34}$$
$$F_k = (N_k(\tau_t) \to \infty \ \text{ as } \ t \to \infty)\,. \tag{C.35}$$

The statement (C.33) implies that $\mathbb{P}(E_k \cup F_k^c) = 1$. If not, suppose $\mathbb{P}(F_k) = 1$, then $0 < \mathbb{P}(E_k^c \cap F_k) = \mathbb{P}(E_k^c)$ which contradicts to the statement $\mathbb{P}(E_k) = 1$. Hence we also have $\mathbb{P}(D) = 1$ where $D := \bigcap_{k\in[K]} E_k \cup F_k^c$.

Now, we prove that, for any random sequence $\{\kappa_{\tau_t} \in [K]\}$,

$$\text{if } \ N_{\kappa_{\tau_t}}(\tau_t) \overset{a.s.}{\to} \infty \ \text{ as } \ t \to \infty \ \text{ then } \ \widehat{\mu}_{\kappa_{\tau_t}}(\tau_t) - \mu_{\kappa_{\tau_t}} \overset{a.s.}{\to} 0 \ \text{ as } \ t \to \infty. \tag{C.36}$$

For notational simplicity, let $Y_k(t) := \widehat{\mu}_k(\tau_t)$, $M_k(t) := N_k(\tau_k)$, and $C_t := \kappa_{\tau_t}$. First, define events G and H such that

$$G = (Y_{C_t}(t) \to \mu_{C_t} \ \text{ as } \ t \to \infty)\,, \tag{C.37}$$
$$H = (M_{C_t}(t)) \to \infty \ \text{ as } \ t \to \infty)\,. \tag{C.38}$$

Note that

$$\begin{aligned}
&|Y_{C_t}(t) - \mu_{C_t}| \to 0, \\
\Leftrightarrow &\sum_{k=1}^{K} \mathbb{1}\left(C_t = k\right) |Y_k(t) - \mu_k| \to 0, \\
\Leftrightarrow &\forall k \in [K], \mathbb{1}\left(C_t = k\right) |Y_k(t) - \mu_k| \to 0, \\
\Leftrightarrow &\forall k \in [K], \mathbb{1}\left(C_t = k\right) \to 0 \ \text{ or } \ |Y_k(t) - \mu_k| \to 0.
\end{aligned}$$

Hence, if $|Y_{C_t}(t) - \mu_{C_t}| \not\to 0$, there exists $k \in [K]$ such that

$$\mathbb{1}(C_t = k) \not\to 0 \ \text{ and } \ |Y_k(t) - \mu_k| \not\to 0.$$

Under the event D, it further implies that there exists $k \in [K]$ such that

$$\mathbb{1}(C_t = k) \not\to 0 \ \text{ and } \ M_k(t) \not\to \infty,$$

which also implies $M_{C_t}(t) \not\to \infty$. Hence, we have $D \cap G^c \subset H^c$ which is equivalent to $H \subset D^c \cup G$. Since $\mathbb{P}(D) = 1$, if $\mathbb{P}(H) = 1$ we have

$$1 = \mathbb{P}(H) \le \mathbb{P}(D^c \cup G) \le \mathbb{P}(D^c) + \mathbb{P}(G) = \mathbb{P}(G),$$

which proves the claimed statement (C.36). From the standard subsequence argument, it also implies that

$$\text{if } \ N_{\kappa_{\tau_t}}(\tau_t) \overset{p}{\to} \infty \ \text{ as } \ t \to \infty \ \text{ if } \ \widehat{\mu}_{\kappa_{\tau_t}}(\tau_t) - \mu_{\kappa_{\tau_t}} \overset{p}{\to} 0 \ \text{ as } \ t \to \infty, \tag{C.39}$$

as desired.

C.5.2 Proof of Proposition 4.4

For any fixed nonadaptive sampling scheme $\nu \in \mathbb{V}$, stopping time $T \in \mathbb{T}$ satisfying $N_k(T) \geq 1$, and for Gaussian arms with mean $\mu_1, \ldots, \mu_K$ and variance σ^2, the likelihood function of given data $\mathcal{D}_T = \{A_1, Y_1, \ldots, A_T, Y_T\}$ with respect to $\mu := (\mu_1, \ldots, \mu_K)$ is proportional to the following expression:

$$\begin{aligned} P(\mathcal{D}_T|\mu) &\propto \prod_{t=1}^{T} \nu_t\left(A_t|\mathcal{D}_{t-1}\right) p_{A_t}\left(Y_t|\mu_{A_t}\right) \\ &\propto \left[\prod_{t=1}^{T} \nu_t\left(A_t|\mathcal{D}_{t-1}\right)\right] \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(Y_t - \mu_{A_t}\right)^2\right\} \\ &= \left[\prod_{t=1}^{T} \nu_t\left(A_t|\mathcal{D}_{t-1}\right)\right] \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{1}\left(A_t = k\right)\left(Y_t - \mu_k\right)^2\right\} \end{aligned}$$

Now, put an independent Gaussian prior with precision $\rho > 0$ to each μ_k, that is, $\pi(\mu_k) \propto \exp\left\{-\frac{\rho}{2}\mu_k^2\right\}$, $\forall k \in [K]$. Then, the posterior distribution of μ can be expressed as follows:

$$\begin{aligned} \pi\left(\mu|\mathcal{D}_T\right) &\propto \left[\prod_{t=1}^{T} \nu_t\left(A_t|\mathcal{D}_{t-1}\right)\right] \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{1}\left(A_t = k\right)\left(Y_t - \mu_k\right)^2\right\} \exp\left\{-\frac{\rho}{2}\sum_{k=1}^{K}\mu_k^2\right\} \\ &\propto \prod_{k=1}^{K} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\mathbb{1}\left(A_t = k\right)\left(\mu_k^2 - 2Y_t\mu_k\right) - \frac{\rho}{2}\mu_k^2\right\} \\ &= \prod_{k=1}^{K} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(N_k(T) + \rho\sigma^2\right)\mu_k^2 - 2N_k(T)\overline{Y}_k(T)\mu_k\right]\right\} \\ &\propto \prod_{k=1}^{K} \exp\left\{-\frac{N_k(T) + \rho\sigma^2}{2\sigma^2}\left(\mu_k - \frac{N_k(T)}{N_k(T) + \rho/\sigma^2}\overline{Y}_k(T)\right)^2\right\}, \end{aligned}$$

where $\overline{Y}_k(T)$ is the sample average of observations from k-th arm. Therefore, the posterior distribution of $(\mu_1, \ldots, \mu_k)$ is coordinate-wisely independent given data and, for each arm, the posterior distribution is given as

$$\mu_k|D_T \sim N\left(\frac{N_k(T)}{N_k(T) + \rho/\sigma^2}\overline{Y}_k(T), \frac{\sigma^2}{N_k(T) + \rho\sigma^2}\right).$$

Hence, for each k, $\widehat{\mu}_k^B := \frac{N_k(T)}{N_k(T)+\rho/\sigma^2}\overline{Y}_k(T)$ is the Bayes estimator for ℓ_2 loss under the Gaussian prior with precision ρ. Denote $P^N|\mu$ be the distribution of the data $\mathcal{D}_T$ under Gaussian arms with mean $\mu = (\mu_1, \ldots, \mu_K)$ and variance σ^2, and let $P^{N,\pi}$ be the posterior predictive distribution under a prior π on

μ. Then, we have the following lower bound on the Bayes risk.

$$\begin{aligned}
\inf_{\widehat{\mu}_k} \mathbb{E}_{\mu\sim\pi}\mathbb{E}_{\mathcal{D}_T\sim P^N|_\mu} N_k(T)\left(\widehat{\mu}_k-\mu_k\right)^2 &= \inf_{\widehat{\mu}_k} \mathbb{E}_{\mathcal{D}_T\sim P^{N,\pi}} N_k(T)\mathbb{E}_{\mu\sim\pi(\cdot|\mathcal{D}_T)}\left(\widehat{\mu}_k-\mu_k\right)^2 \\
&\geq \mathbb{E}_{\mathcal{D}_T\sim P^{N,\pi}} N_k(T) \inf_{\widehat{\mu}_k} \mathbb{E}_{\mu\sim\pi(\cdot|\mathcal{D}_T)}\left(\widehat{\mu}_k-\mu_k\right)^2 \\
&= \mathbb{E}_{\mathcal{D}_T\sim P^{N,\pi}} N_k(T)\mathbb{E}_{\mu\sim\pi(\cdot|\mathcal{D}_T)}\left(\widehat{\mu}_k^B-\mu_k\right)^2 \\
&\geq \mathbb{E}_{\mathcal{D}_T\sim P^{N,\pi}}\left[\frac{\sigma^2 N_k(T)}{N_k(T)+\rho\sigma^2}\right] \\
&= \mathbb{E}_{\mathcal{D}_T\sim P^{N,\pi}}\left[\frac{\sigma^2}{1+\rho\sigma^2/N_k(T)}\right],
\end{aligned}$$

where the last equality comes from the assumption $N_k(T)\geq 1$.

Based on the Bayes risk calculation above, we can find a lower bound on the minimax normalized ℓ_2 risk for each $\rho>0$ as follows:

$$\begin{aligned}
\inf_{\widehat{\mu}_k}\sup_{\substack{P_k\in\mathbb{P}_k(\mu_k,\sigma_k)\\ \nu\in\mathbb{V},T\in\mathbb{T}}} \mathbb{E}_Q N_k(T)\left(\widehat{\mu}_k-\mu_k\right)^2 &\geq \inf_{\widehat{\mu}_k} \mathbb{E}_{\mu\sim\pi}\mathbb{E}_{\mathcal{D}_T\sim P^N|_\mu} N_k(T)\left(\widehat{\mu}_k-\mu_k\right)^2 \\
&\geq \mathbb{E}_{\mathcal{D}_T\sim P^{N,\pi}}\left[\frac{\sigma^2}{1+\rho\sigma^2/N_k(T)}\right].
\end{aligned}$$

Since we assume sampling and stopping strategies are nonadaptive, the distribution of $N_k(T)$ does not depend on π. Therefore, by the monotone convergence theorem with $\rho\searrow 0$, we have the following lower bound on the minimax normalized ℓ_2 risk.

$$\inf_{\widehat{\mu}_k}\sup_{\substack{P_k\in\mathbb{P}_k(\mu_k,\sigma_k)\\ \nu\in\mathbb{V},T\in\mathbb{T}}} \mathbb{E}_Q N_k(T)\left(\widehat{\mu}_k-\mu_k\right)^2 \geq \sigma^2. \tag{C.40}$$

From the nonadaptivity of data collecting procedure, it can be easily shown that, for any choice of $P_k\in\mathbb{P}_k(\mu_k,\sigma_k), \nu\in\mathbb{V}, T\in\mathbb{T}$ and the corresponding $Q=Q(P_k,\nu,T)$, we have

$$\mathbb{E}_Q N_k(T)\left(\overline{Y}_k(T)-\mu_k\right)^2=\sigma_k^2,$$

which shows that the minimax risk is equal to σ_k^2 and the sample mean estimator achieves it as claimed.

C.5.3 Proof of Proposition 4.5

The proof of Proposition 4.5 relies on a lower bound of D_{f_q} and arguments in Jiao et al. [2017] which is summarized in Lemma C.7 in Section C.2.2.

To apply the lemma, we first prove the following bound on the expectation of the p-norm of the normalized ℓ_2 loss.

Claim C.13. *Under the assumptions of Proposition 4.5, for each $k\in[K]$ and for any fixed $p>1$ we have that*

$$\left\|N_k(T)\left(\widehat{\mu}_k(T)-\mu_k\right)^2\right\|_p \leq C_p\left(\sigma_k^{(2p)}\right)^2, \tag{C.41}$$

where C_p is a constant depending only on p.

The proof of the claim is based on the Marcinkiewicz-Zygmund (M-Z) inequality. We cite the following form of the inequality for completeness.

Lemma C.14 (Marcinkiewicz and Zygmund [1937]). *For any $p \geq 1$, if $X_1, \dots X_n$ are independent random variables with $\mathbb{E}[X_i] = 0$ and $\mathbb{E}|X_i|^p < \infty$ for all $i = 1, \dots, n$ then the following inequality holds.*

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right|^p\right] \leq B_p \mathbb{E}\left[\left(\sum_{i=1}^{n} |X_i|^2\right)^{p/2}\right], \tag{C.42}$$

where $B_p > 0$ is a constant depending only on p.

Proof of Claim C.13. For simple notations, let $W_t := \mathbb{1}(A_t = k)$ and $Z_t = Y_t - \mu_k$. Then from the M-Z inequality, we have

$$\begin{aligned}
&\mathbb{E}\left|N_k(T)\left(\widehat{\mu}_k(T) - \mu_k\right)^2\right|^p \\
&= \mathbb{E}\left|\frac{1}{\sqrt{N_k(T)}} \sum_{t=1}^{T} W_t Z_t\right|^{2p} \\
&= \mathbb{E}\left[\mathbb{E}\left[\left|\frac{1}{\sqrt{N_k(T)}} \sum_{t=1}^{T} W_t Z_t\right|^{2p} \mid \{W_t\}_{t\geq 1}\right]\right] \\
&\leq B_p \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{N_k(T)} \sum_{t=1}^{T} W_t |Z_t|^2\right)^p \mid \{W_t\}_{t\geq 1}\right]\right] \quad \text{(by M-Z inequality and } W_t^2 = W_t) \\
&\leq B_p \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{N_k(T)} \sum_{t=1}^{T} W_t |Z_t|^{2p}\right) \mid \{W_t\}_{t\geq 1}\right]\right] \quad \text{(by Jensen's inequality with } N_k(T) = \sum_{t=1}^{T} W_t) \\
&\leq B_p \left(\sigma^{(2p)}\right)^{2p},
\end{aligned}$$

which implies the claimed inequality with $C_p := (B_p)^{1/p}$. □

Now, we have all building blocks to complete the proof of Proposition 4.5. For each $k \in [K]$, set $P_k = \mathcal{L}(\mathcal{D}_T | \kappa = k)$, $Q = \mathcal{L}(\mathcal{D}_T)$ and

$$f_k = \lambda N_k(T)\left(\widehat{\mu}_k(T) - \mu_k\right)^2.$$

for a $\lambda > 0$. Then, from Lemma C.7, we can lower bound $I_q(\kappa, \mathcal{D}_T)$ in the following way:

$$\begin{aligned}
\frac{1}{q} I_q(\kappa, \mathcal{D}_T) &= \sum_{k=1}^{K} \mathbb{P}(\kappa = k) \left\{ \frac{1}{q} D_{f_q}\left(\mathcal{L}(\mathcal{D}_T|\kappa = k) \,||\, \mathcal{L}(\mathcal{D}_T)\right) \right\} \\
&\geq \sum_{k=1}^{K} \mathbb{P}(\kappa = k) \left\{ \mathbb{E}_{P_k}[f_k] - \mathbb{E}_Q[f_k] - \mathbb{E}_Q\left[\frac{|f_k|^p}{p}\right] \right\} \\
&= \sum_{k=1}^{K} \mathbb{P}(\kappa = k) \left\{ \lambda \mathbb{E}\left[N_k(T)\left(\widehat{\mu}_k(T) - \mu_k\right)^2 \mid \kappa = k\right] \right. \\
&\qquad \left. -\lambda \mathbb{E}\left[N_k(T)\left(\widehat{\mu}_k(T) - \mu_k\right)^2\right] - \frac{\lambda^p}{p} \mathbb{E}\left[N_k(T)\left(\widehat{\mu}_k(T) - \mu_k\right)^2\right]^p \right\} \\
&\geq \sum_{k=1}^{K} \mathbb{P}(\kappa = k) \left\{ \lambda \mathbb{E}\left[N_k(T)\left(\widehat{\mu}_k(T) - \mu_k\right)^2 \mid \kappa = k\right] - \left(\lambda \sigma_k^2 + (\lambda C_p (\sigma_k^{(2p)})^2)^p / p\right) \right\} \\
&= \lambda \mathbb{E}\left[N_\kappa(T)\left(\widehat{\mu}_\kappa(T) - \mu_\kappa\right)^2\right] - \left(\lambda \|\sigma_\kappa\|_2^2 + \frac{\lambda^p C_p^p}{p} \|\sigma_\kappa^{(2p)}\|_{2p}^{2p}\right).
\end{aligned}$$

Since this inequality holds for any $\lambda > 0$, we get

$$\begin{aligned}
\mathbb{E}\left[N_\kappa(T)\left(\widehat{\mu}_\kappa(T) - \mu_\kappa\right)^2\right] &= \|\sigma_\kappa\|_2^2 + \inf_{\lambda > 0} \frac{1}{\lambda} \left\{ \frac{I_q(\kappa, \mathcal{D}_T)}{q} + \frac{\lambda^p C_p^p}{p} \|\sigma_\kappa^{(2p)}\|_{2p}^{2p} \right\} \\
&= \|\sigma_\kappa\|_2^2 + C_p \|\sigma_\kappa^{(2p)}\|_{2p}^2 I_q^{1/q}(\kappa, \mathcal{D}_T),
\end{aligned}$$

which completes the proof.

C.5.4 Proof of Fact 4.10

Let θ_1 and θ_0 be natural parameters corresponding to μ_1 and μ_0 such that $\mu_1 = B'(\theta_1)$ and $\mu_0 = B'(\theta_0)$. From well-known fact about the KL divergence in exponential family theory,

$$\begin{aligned}
\ell_{KL}(\mu_1, \mu_0) &= D_{KL}(\theta_1 \| \theta_0) \\
&= B'(\theta_1)(\theta_1 - \theta_0) - B(\theta_1) + B(\theta_0).
\end{aligned}$$

Since $\mu = B'(\theta)$ and $\psi_\mu(\lambda) := \lambda\mu + \psi(\lambda; \theta) = B(\lambda + \theta) - B(\theta)$, its derivative with respect to λ is equal to $\psi'_\mu(\lambda) = B'(\lambda + \theta)$. Thus,

$$\begin{aligned}
D_{KL}(\theta_1 \| \theta_0) &= B'(\theta_1)(\theta_1 - \theta_0) - B(\theta_1) + B(\theta_0) \\
&= \psi'_{\mu_0}(\theta_1 - \theta_0)(\theta_1 - \theta_0) - \psi_{\mu_0}(\theta_1 - \theta_0) \\
&= \psi_{\mu_0}(0) - \psi_{\mu_0}(\theta_1 - \theta_0) - \psi'_{\mu_0}(\theta_1 - \theta_0)(0 - (\theta_1 - \theta_0)) \quad (\text{since } \psi_{\mu_0}(0) = 0.) \\
&= D_{\psi_{\mu_0}}(0, \theta_1 - \theta_0) \\
&= D_{\psi^*_{\mu_0}}\left(\psi'_{\mu_0}(\theta_1 - \theta_0), \mu_0\right) \\
&= D_{\psi^*_{\mu_0}}(\mu_1, \mu_0),
\end{aligned}$$

where the last equality comes from the fact that $\psi'_{\mu_0}(\theta_1 - \theta_0) = B'(\theta_1) = \mu_1$. The second-to-last equality stems instead from the duality of the Bregman divergence, which we state below.

Fact C.15. *Let $f : \Lambda \to \mathbb{R}$ be a strictly convex and continuously differentiable on a open interval $\Lambda \subset \mathbb{R}$. For any $\theta_1, \theta_0 \in \Lambda$, let μ_1, μ_2 be the corresponding dual points satisfying $f'(\theta_j) = \mu_j$ for $j = 0, 1$. Then, we have*

$$D_f(\theta_0, \theta_1) = D_{f^*}(\mu_1, \mu_0), \tag{C.43}$$

where f^* is the convex conjugate of f.

Therefore, we have the first claimed equality, $\ell_{KL}(\mu_1, \mu_0) = D_{\psi^*_{\mu_0}}(\mu_1, \mu_0)$. To show the second one, first note that $\psi'_{\mu_0}(0) = B'(\theta_0) = \mu_0$, so that $\psi^*_{\mu_0}(\mu_0) = \psi^{*\prime}_{\mu_0}(\mu_0) = 0$. Therefore, we have that

$$D_{\psi^*_{\mu_0}}(\mu_1, \mu_0) = \psi^*_{\mu_0}(\mu_1) - \psi^*_{\mu_0}(\mu_0) - \psi^{*\prime}_{\mu_0}(\mu_0)\,(\mu_1 - \mu_0) = \psi^*_{\mu_0}(\mu_1),$$

which verifies the second claimed equality. The last equality can be established as follows:

$$\begin{aligned}
\psi^*_{\mu_0}(\mu_1) &= \sup_{\lambda} \lambda\mu_1 - \psi_{\mu_0}(\lambda) \\
&= \sup_{\lambda} \lambda\mu_1 - [\lambda\mu_0 + \psi(\lambda)] \\
&= \sup_{\lambda} \lambda\,(\mu_1 - \mu_0) - \psi(\lambda) \\
&= \psi^*(\mu_1 - \mu_0),
\end{aligned}$$

as desired.

C.5.5 Proof of Proposition 4.11

By following similar arguments to the ones used in the proof of Lemma 4.13, we first show that, for any $\delta > 0$, the following deviation inequality holds.

$$\mathbb{P}\left(D_{\psi^*_\mu}(\widehat{\mu}(n), \mu) \geq \delta\right) \leq 2e^{-n\delta}. \tag{C.44}$$

Proof of inequality (C.44). For any $\epsilon \geq 0$ and $\lambda \in [0, \lambda_{\max}) \subset \Lambda$, we have

$$\begin{aligned}
\mathbb{P}\left(S_k(n)/n - \mu_k \geq \epsilon\right) &= \mathbb{P}\left(S_k(n) - n\mu_k \geq n\epsilon\right) \\
&= \mathbb{P}\left(e^{\lambda(S_k(n) - n\mu_k)} \geq e^{\lambda n\epsilon}\right) \\
&\leq e^{-\lambda n\epsilon}\mathbb{E}\left[e^{\lambda(S_k(n) - n\mu_k)}\right],
\end{aligned}$$

where in the final step we have used Markov's inequality. The last term can be bounded as follows:

$$\begin{aligned}
e^{-\lambda n\epsilon}\mathbb{E}\left[e^{\lambda(S_k(n) - n\mu_k)}\right] &= e^{-\lambda n\epsilon}\prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda(X_i - \mu_k)}\right] \\
&= e^{-\lambda n\epsilon}\prod_{i=1}^{n}\mathbb{E}\left[e^{\psi(\lambda)}\right] \\
&= e^{n(\psi(\lambda) - \lambda\epsilon)}.
\end{aligned}$$

Since the bound holds for any $\lambda \in [0, \lambda_{\max})$, we have the following intermediate bound on the deviation probability:

$$\mathbb{P}\left(S_k(n)/n - \mu_k \geq \epsilon\right) \leq \inf_{\lambda \in [0, \lambda_{\max})} e^{n(\psi(\lambda) - \lambda\epsilon)}. \tag{C.45}$$

Since $\epsilon \geq 0$, the convex conjugate of ψ at ϵ can be written as

$$\psi^*(\epsilon) = \sup_{\lambda \in \Lambda} \{\lambda\epsilon - \psi(\lambda)\} = \sup_{\lambda \in [0, \lambda_{\max})} \{\lambda\epsilon - \psi(\lambda)\}.$$

Using this identity, the deviation probability can be further bounded as

$$\begin{aligned}
\mathbb{P}\left(S_k(n)/n - \mu_k \geq \epsilon\right) &\leq \inf_{\lambda \in [0, \lambda_{\max})} e^{n(\psi(\lambda) - \lambda\epsilon)} \\
&= \exp\left(-n \sup_{\lambda \in [0, \lambda_{\max})} \{\lambda\epsilon - \psi(\lambda)\}\right) \\
&= e^{-n\psi^*(\epsilon)}
\end{aligned}$$

Using the same argument, it also follows that

$$\mathbb{P}\left(S_k(n)/n - \mu_k \leq -\epsilon\right) \leq e^{-n\psi^*(-\epsilon)}$$

Since ψ^* is a non-negative convex function with $\psi^*(0) = 0$, for any $\delta \geq 0$, there exist $\epsilon_1, \epsilon_2 \geq 0$ with $\psi^*(\epsilon_1) = \psi^*(-\epsilon_2) = \delta$ such that

$$\{z \in \mathbb{R} : \psi^*(z) \geq \delta\} = \{z \in \mathbb{R} : z \geq \mu_k + \epsilon_1, z \leq \mu_k - \epsilon_2\}.$$

Therefore, for any $\delta \geq 0$, we conclude that

$$\begin{aligned}
\mathbb{P}\left(D_{\psi^*_{\mu_k}}(\widehat{\mu}_k(n), \mu_k) \geq \delta\right) &= \mathbb{P}\left(\psi^*_{\mu_k}(S_k(n)/n) \geq \delta\right) \quad \text{(By the equality (4.25) in Fact 4.10.)} \\
&\leq \mathbb{P}\left(S_k(n)/n) - \mu_k \geq \epsilon_1\right) + \mathbb{P}\left(S_k(n)/n) - \mu_k \leq -\epsilon_2\right) \\
&\leq 2e^{-n\delta},
\end{aligned}$$

as desired. □

Now, we return to the proof of Proposition 4.11. Based on the deviation inequality (C.44), the risk under the non-adaptive setting can be bounded as

$$\begin{aligned}
\mathbb{E}_\theta D_{\psi^*_\mu}(\widehat{\mu}(n), \mu) &= \int_0^\infty \mathbb{P}\left(D_{\psi^*_\mu}(\widehat{\mu}(n), \mu) \geq \delta\right) \mathrm{d}\delta \\
&\leq 2\int_0^\infty e^{-n\delta} \mathrm{d}\delta \\
&= \frac{2}{n},
\end{aligned}$$

which completes the proof of the upper bound. To get a lower bound on the minimax risk, we use the following lemma.

Lemma C.16 (Modified version of Theorem 2.2 in Tsybakov [2008]). *Let $\{P_\theta : \theta \in \Theta\}$ be a family of probability measures parameterized by $\theta \in \Theta$ and let $s > 0$. Suppose a loss function $l : \Theta \times \Theta \mapsto [0, \infty)$ satisfies the local triangle inequality condition with positive numbers $M \leq 1$ and ϵ_0. Also assume that there exist $\theta_1, \theta_0 \in \Theta$ such that $\ell(\theta_1, \theta_0) \geq 2s$ for some $s \leq \epsilon_0$. Then, if $D_{KL}(P_{\theta_1} \| P_{\theta_0}) \leq \alpha < \infty$, we have*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta\left(\ell(\widehat{\theta}, \theta) \geq s\right) \geq \frac{M}{4}\exp(-\alpha). \tag{C.46}$$

Note that, for any $\theta_1, \theta_0 \in \Theta$, the KL divergence can be written as

$$D_{KL}(P^n_{\theta_1} || P^n_{\theta_0}) = nD_{KL}(\theta_1 || \theta_0) = nD_{\psi^*_{\mu_0}}(\mu_1, \mu_0),$$

where μ_1 and μ_0 are corresponding mean parameters. If n is large enough such that $\frac{\log 2}{2n} \leq \epsilon_0$, we can always find $\theta_1, \theta_0 \in \Theta$ such that $D_{\psi^*_{\mu_0}}(\mu_1, \mu_0) = \frac{\log 2}{n}$. Then, the condition in Lemma C.16 can be satisfied with $\ell = D_{\psi^*_\mu}$, $s = \frac{\log 2}{2n}$ and $\alpha = \log 2$. Therefore,

$$\begin{aligned}
\inf_{\widehat{\mu}} \sup_{\mu} \mathbb{E}_\theta D_{\psi^*_\mu}(\widehat{\mu}, \mu) &\geq s \inf_{\widehat{\mu}} \sup_{\mu} \mathbb{P}\left(D_{\psi^*_\mu}(\widehat{\mu}, \mu) \geq s \right) \\
&\geq \frac{sM}{4} \exp(-\alpha) \\
&= \frac{M \log 2}{16n},
\end{aligned}$$

as desired.

C.6 Equivalence between $n_t^{\text{eff}} \to \infty$ and $N(t) \xrightarrow{a.s.} \infty$

Before we state and formally prove our claim, we first state a useful fact.

Fact C.17 (Theorem 13.7 in Williams [1991] with $X = 0$)**.** *Let $\{X_t\}_{t\in\mathbb{N}}$ be a sequence of random variables with finite first moments. Then $\mathbb{E}|X_t| \to 0$ if and only if the following conditions are satisfied:*

1. $X_t \xrightarrow{p} 0$.
2. $\{X_t\}_{t\in\mathbb{N}}$ *is uniformly integrable.*

Recall that $n_t^{\text{eff}} = [1/N(t)]^{-1}$, we are now in place to prove the following claim.

Proposition C.18. *As long as $N(t) \geq b > 0$, we have that $n_t^{\text{eff}} \to \infty$ as $t \to \infty$ if and only if $N(t) \xrightarrow{p} \infty$ $t \to \infty$.*

Proof. The assumption of $N(t) \geq b > 0$ ensures that the sequence $\{1/N(t)\}_{t\in\mathbb{N}}$ is uniformly integrable. Substituting $X_t = 1/N(t)$ into the aforementioned fact, we have that

$$\begin{aligned}
n_t^{\text{eff}} \to \infty \text{ as } t \to \infty &\Leftrightarrow \mathbb{E}\left[1/N(t)\right] \to 0 \text{ as } t \to \infty \\
&\overset{\text{fact}}{\Leftrightarrow} 1/N(t) \xrightarrow{p} 0 \text{ as } t \to \infty \\
&\Leftrightarrow N(t) \xrightarrow{p} \infty \text{ as } t \to \infty,
\end{aligned}$$

as desired. □

Proposition C.19. *If $\{N(t)\}$ is a nondecreasing sequence, we have that $N(t) \xrightarrow{p} \infty$ as $t \to \infty$ implies $N(t) \xrightarrow{a.s.} \infty$ as $t \to \infty$.*

Proof. If $N(t) \xrightarrow{p} \infty$ as $t \to \infty$, there exists a subsequence goes to ∞ almost surely. Therefore, we must have $\limsup_{t\to\infty} N(t) = \infty$ almost surely. By the monotonicity of $\{N(t)\}$, we have $\limsup_{t\to\infty} N(t) = \lim_{t\to\infty} N(t)$ which implies that $N(t) \xrightarrow{a.s.} \infty$ as $t \to \infty$. □

Returning to the MABs setting, the previous propositions show that, as long as $N_k(t) \geq b > 0$, the condition $n_{k,t}^{\text{eff}} \to \infty$ implies that $N_k(t) \xrightarrow{a.s.} \infty$ since $\{N_k(t)\}$ is monotone for each arm $k \in [K]$. For a sequence of chosen arms, however, if the sequence $\{N_{\kappa_t}(T_t)\}$ is not monotone, $n_{\kappa_t,t}^{\text{eff}} \to \infty$ does not imply $N_{\kappa_t}(\tau_t) \xrightarrow{a.s.} \infty$ as shown in the next example.

Example C.20. *Consider a two-armed bandit; pull the first arm at time 1 and the second arm forever after. Thus, $N_1(t) = 1$ for all $t \geq 1$, and $N_2(t) = t - 1$ for all $t \geq 2$ with $N_2(1) = 0$. Define $\tau_t = t + 1$ and let $\{\kappa_t\}$ be a sequence of random choice functions defined by a uniform random variable $U \in Unif[0, 1]$ such that $\kappa_t = 1$ if $U \in \left[\frac{j}{2^k}, \frac{j+1}{2^k}\right]$ where k and j are given by $k = \lfloor \log_2(t) \rfloor$ and $t = 2^k + j$. if $U \notin \left[\frac{j}{2^k}, \frac{j+1}{2^k}\right]$, define $\kappa_t = 2$. It is clear $N_{\kappa_t}(\tau_t) \xrightarrow{p} \infty$, and the Proposition C.18 implies that $n^{\text{eff}}_{\kappa_t,t} \to \infty$. However, for any given U, $N_{\kappa_t}(\tau_t) \not\to \infty$ and thus $\mathbb{P}\left(N_{\kappa_t}(\tau_t) \to \infty\right) = 0$.*

C.7 Alternative bounds using sub-Gaussian self-normalized process

For sub-Gaussian arms, it is known that $\mathbb{E}\left[\exp\left\{\lambda\left(S(\mathcal{T}) - \mu N(\mathcal{T})\right) - \frac{\sigma^2\lambda^2}{2}N(\mathcal{T})\right\}\right] \leq 1$ for all $\lambda \in \mathbb{R}$. In this sub-Gaussian case (only), one may use the following moment bound from the literature on self-normalized processes.

Fact C.21 (Theorem 2.1 in de la Pena et al. [2004]). *If $\mathbb{E}\sqrt{N(\mathcal{T})} < \infty$,*

$$\mathbb{E}\exp\left\{\widetilde{N}^{\mathrm{E}}(\mathcal{T})\frac{(\widehat{\mu}(\mathcal{T}) - \mu))^2}{4\sigma^2}\right\} \leq \sqrt{2}, \tag{C.47}$$

where $\widetilde{N}^{\mathrm{E}}(\mathcal{T}) := N^2(\mathcal{T}) / \left(N(\mathcal{T}) + \left(\mathbb{E}\sqrt{N(\mathcal{T})}]\right)^2\right)$.

We can use the above fact and the Donsker-Varadhan representation to derive an alternative bound for the l_2 risk of the chosen sample mean at a stopping time $\mathcal{T}$ as follows:

$$\begin{aligned}
I(\kappa; \mathcal{D}) &= \sum_{k=1}^{K} \mathbb{P}(\kappa = k) D_{KL}\left(\mathcal{L}\left(\mathcal{D}|\kappa = k\right) || \mathcal{L}\left(\mathcal{D}\right)\right) \\
&\geq \sum_{k=1}^{K} \mathbb{P}(\kappa = k)\mathbb{E}_{P_k}\left[f_k\right] - \log \mathbb{E}_Q\left[e^{f_k}\right] \\
&= \sum_{k=1}^{K} \mathbb{P}(\kappa = k)\left\{\mathbb{E}\left[\widetilde{N}^{\mathrm{E}}_k(\mathcal{T})\frac{(\widehat{\mu}_k(\mathcal{T}) - \mu_k))^2}{4\sigma^2} \mid \kappa = k\right]\right. \\
&\qquad \left. - \log \mathbb{E}\left[\exp\left\{\widetilde{N}^{\mathrm{E}}_k(\mathcal{T})\frac{(\widehat{\mu}(\mathcal{T}) - \mu))^2}{4\sigma^2}\right\}\right]\right\} \\
&\geq \sum_{k=1}^{K} \mathbb{P}(\kappa = k)\left\{\mathbb{E}\left[\widetilde{N}^{\mathrm{E}}_k(\mathcal{T})\frac{(\widehat{\mu}_k(\mathcal{T}) - \mu_k))^2}{4\sigma^2} \mid \kappa = k\right] - \frac{\log 2}{2}\right\} \\
&= \mathbb{E}\left[\widetilde{N}^{\mathrm{E}}_\kappa(\mathcal{T})\frac{(\widehat{\mu}_\kappa(\mathcal{T}) - \mu_\kappa))^2}{4\sigma^2}\right] - \frac{\log 2}{2}.
\end{aligned}$$

By rearranging terms, we have the following bound on the ℓ_2 risk.

$$\mathbb{E}\left[\widetilde{N}^{\mathrm{E}}_\kappa(\mathcal{T})\left(\widehat{\mu}_\kappa(\mathcal{T}) - \mu_\kappa)\right)^2\right] \leq 4\sigma^2\left[I(\kappa; \mathcal{D}_\mathcal{T}) + \frac{\log 2}{2}\right]. \tag{C.48}$$

Recall that, for sub-Gaussian arms, the bound in Theorem 4.16 can be written as

$$\mathbb{E}\left[\widetilde{\widetilde{N}}_\kappa(\tau)\left(\widehat{\mu}_\kappa(\tau) - \mu_\kappa)\right)^2\right] \leq 2C_b\sigma^2\left[I(\kappa; \mathcal{D}_\mathcal{T}) + 1.25\right]. \tag{C.49}$$

Bounds (C.48) and (C.49) are matched to each other up to a constant factor. However, corresponding normalized ℓ_2 risks in LHS shows interesting differences. First, the bound (C.48) based on Fact C.21 holds only at a stopping time but our bound holds at an arbitrary random time. Second, the bound (C.48) is applicable only to the sub-Gaussian case since it is non-trivial to extend the Fact C.21 to general sub-ψ cases. Third, if the random sample size N is highly concentrated at a constant, the normalizing factor $\widetilde{N}_\kappa^{\mathrm{E}}$ tends to be larger than our normalizing factor $\widetilde{\widetilde{N}}_\kappa$ and thus the bound (C.48) yields a tighter control on the ℓ_2 risk. On the other hand, if the random sample size N has a large variability, our normalizing factor $\widetilde{\widetilde{N}}_\kappa$ tends to be larger than $\widetilde{N}_\kappa^{\mathrm{E}}$ since $(\mathbb{E}\sqrt{N})^2$ can be significantly larger than N with a high probability. In this case, our bound (C.49) yields a tighter control on the ℓ_2 risk than the bound (C.48).

Bibliography

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012. 1.1.1

Francis J Anscombe. Large-sample theory of sequential estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 600–607. Cambridge University Press, 1952. 4.1

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009. 1.1.1

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002. 1.1.1

Akshay Balsubramani. Sharp Finite-Time Iterated-Logarithm Martingale Concentration. *arXiv:1405.2639 [cs, math, stat]*, May 2014. 4.1

Akshay Balsubramani and Aaditya Ramdas. Sequential Nonparametric Testing with the Law of the Iterated Logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 42–51, Arlington, Virginia, 2016. AUAI Press. ISBN 978-0-9966431-1-5. 4.1

Jack Bowden and Lorenzo Trippa. Unbiased estimation for response adaptive clinical trials. *Statistical methods in medical research*, 26(5):2376–2388, 2017. 2, 2.1.2

María Emilia Caballero, Begoña Fernández, and David Nualart. Estimation of densities and applications. *Journal of Theoretical Probability*, 11(3):831–851, 1998. 4.4.2

DR Cox. A note on the sequential estimation of means. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 447–450. Cambridge University Press, 1952. 4.1

D. A. Darling and Herbert Robbins. Confidence Sequences for Mean, Variance, and Median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, July 1967a. ISSN 0027-8424, 1091-6490. 4.1

D. A. Darling and Herbert Robbins. Inequalities for the Sequence of Sample Means. *Proceedings of the National Academy of Sciences*, 57(6):1577–1580, June 1967b. ISSN 0027-8424. C.2.1

D. A. Darling and Herbert Robbins. Some Further Remarks on Inequalities for Sample Sums. *Proceedings of the National Academy of Sciences*, 60(4):1175–1182, August 1968. ISSN 0027-8424. 4.1

Victor H de la Pena, Michael J Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of probability*, pages 1902–1933, 2004. 4.1, C.21

Yash Deshpande, Lester Mackey, Vasilis Syrgkanis, and Matt Taddy. Accurate inference for adaptive linear models. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. 3.3, 4.6

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983. C.4.2

Aurélien Garivier. Informational confidence bounds for self-normalized averages and applications. In *2013 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2013. 4.4.3, C.4.1

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference On Learning Theory*, pages 359–376, 2011. 1.1.1, C.2.1, C.3.1, C.4.1

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027, 2016. 2.2.3

Allan Gut. *Stopped random walks*. Springer, 2009. 2, 3.1, 4.1, 4.2, 4.2, 4.2, C.5.1

Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *arXiv preprint arXiv:1911.02768*, 2019. 3.3

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Exponential line-crossing inequalities. *arXiv preprint arXiv:1808.03204*, 2018a. 4.5, C.3.1

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018b. 2.2.2, 2.2.2, 4.1, C.2.1, C.4.1

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' UCB: An Optimal Exploration Algorithm for Multi-Armed Bandits. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 423–439, 2014. 1.1.1, 2.1.2, 2.2.3, 3.2.3, 4.1

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Dependence measures bounding the exploration bias for general measurements. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 1475–1479. IEEE, 2017. 4.1, 4.3.1, 4.3.1, 4.3.1, 4.3.2, 4.5, 4.6, C.2.2, C.4.2, C.5.3

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012. 1.1.1

Emilie Kaufmann and Wouter Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *arXiv preprint arXiv:1811.11419*, 2018. 4.4.3

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012. 1.1.1

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016. 4.1

Rasul A Khan. L_p-version of the Dubins–Savage inequality and some exponential inequalities. *Journal of Theoretical Probability*, 22(2):348, 2009. 4.3.2, 4.5, C.2.1, C.6

Tze Leung Lai. On Confidence Sequences. *The Annals of Statistics*, 4(2):265–280, March 1976. ISSN 0090-5364, 2168-8966. 4.1

Lihong Li, Remi Munos, and Csaba Szepesvari. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pages 608–616, 2015. 1

Józef Marcinkiewicz and Antoni Zygmund. Sur les fonctions indépendantes. *Fundamenta Mathematicae*, 29(1):60–90, 1937. 4.3.1, C.14

Seth Neel and Aaron Roth. Mitigating bias in adaptive data gathering via differential privacy. *arXiv preprint arXiv:1806.02329*, 2018. 3.3, 4.6

Xinkun Nie, Xiaoying Tian, Jonathan Taylor, and James Zou. Why adaptively collected data have negative bias and how to correct for it. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269, 2018. 2, 2, 2.1.1, 2.1, 2.1.2, 2.7, 2.1.2, 2.1.2, 3.3, 4.6, A.1.1, A.1.1, A.1.1, A.1.1, A.1.3

Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008. 4.1, 4.4.2, C.3.2

Wolfgang Richter. Limit theorems for sequences of random variables with sequences of random indeces. *Theory of Probability & Its Applications*, 10(1):74–84, 1965. 4.1

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952. 1

Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970. 2.2.2

Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418, 2016. 2.2.3

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240, 2016. 4.1, 4.4.3, 4.4.3, 4.4.3, 4.5, 4.6, C.4.2

Harold Sackrowitz and Ester Samuel-Cahn. Evaluating the chosen population: a Bayes and minimax approach. *Lecture Notes-Monograph Series*, pages 386–399, 1986. 4.6

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. On the bias, risk and consistency of sample means in multi-armed bandits. *arXiv preprint arXiv:1902.00746*, 2019a. 1, 3, A.2.1, B.1.2

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. Are sample means in multi-armed bandits positively or negatively biased? In *Advances in Neural Information Processing Systems*, pages 7100–7109, 2019b. 1

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. On conditional versus marginal bias in multi-armed bandits. *arXiv preprint arXiv:2002.08422*, 2020. 1

David Siegmund. Estimation following sequential tests. *Biometrika*, 65(2):341–349, 1978. 2.1.1, 4.1

David Siegmund. *Sequential Analysis: Tests and Confidence Intervals*. Springer Science & Business Media, 1985. 4.4.2

Norman Starr. On the asymptotic efficiency of a sequential procedure for estimating the mean. *The Annals of Mathematical Statistics*, 37(5):1173–1185, 1966. 4.1

Norman Starr and Michael Woodroofe. Further remarks on sequential estimation: the exponential case. *The Annals of Mathematical Statistics*, pages 1147–1154, 1972. 4.1

Norman Starr and Michael B Woodroofe. Remarks on a stopping time. *Proceedings of the National Academy of Sciences of the United States of America*, 61(4):1215, 1968. 2, 2.1.1, 2.5, 3

Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*. MIT press Cambridge, 1998. 1.1.1

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. 1.1.1

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510. C.16

Sofía S Villar, Jack Bowden, and James 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, 30(2):199, 2015. 2

J Ville. Étude critique de la notion de collectif. *Gauthier-Villars, Paris*, 1939. C.4.1

Abraham Wald and Jacob Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, pages 326–339, 1948. 4.1

David Williams. *Probability with martingales*. Cambridge university press, 1991. C.17

Min Xu, Tao Qin, and Tie-Yan Liu. Estimation bias in multi-armed bandit algorithms for search advertising. In *Advances in Neural Information Processing Systems*, pages 2400–2408, 2013. 2, 3.3, 4.6, A.1.3

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999. 4.4.1