Adaptive Sequential Experiments with Unknown Information Flows

Yonatan Gur
Stanford University

Ahmadreza Momeni*
Stanford University

February 7, 2019

Abstract

An agent facing sequential decisions that are characterized by partial feedback needs to strike a balance between maximizing immediate payoffs based on available information, and acquiring new information that may be essential for maximizing future payoffs. This trade-off is captured by the multi-armed bandit (MAB) framework that has been studied and applied under strong assumptions on the information collection process: at each time epoch a single observation is collected on the action that was selected at that epoch. However, in many practical settings additional information may become available between pulls, and might be essential for achieving good performance. We introduce a generalized MAB formulation that relaxes the strong assumptions on the information collection process, and in which auxiliary information on each arm may appear arbitrarily over time. By obtaining matching lower and upper bounds, we characterize the (regret) complexity of this family of MAB problems as a function of the information flows, and study how salient characteristics of the information impact policy design and achievable performance. We introduce a broad adaptive exploration approach for designing policies that, without any prior knowledge on the information arrival process, attain the best performance (regret rate) that is achievable when the information arrival process is a priori known. Our approach is based on adjusting MAB policies designed to perform well in the absence of auxiliary information by using dynamically customized virtual time indexes to endogenously control the exploration rate of the policy. We demonstrate the effectiveness of our approach through establishing performance bounds and evaluating numerically the performance of adjusted well-known MAB policies. Our study demonstrates how decision-making policies designed to perform well with very little information can be adjusted to also guarantee optimality in more information-abundant settings.

Keywords: sequential decisions, data-driven decisions, online learning, adaptive algorithms, multi-armed bandits, exploration-exploitation, minimax regret.

1 Introduction

1.1 Background and motivation

In the presence of uncertainty and partial feedback on payoffs, an agent that faces a sequence of decisions needs to strike a balance between maximizing instantaneous performance indicators (such as revenue) and collecting valuable information that is essential for optimizing future decisions. A well-studied
framework that captures this trade-off between new information acquisition (exploration), and optimizing payoffs based on available information (exploitation) is the one of multi-armed bandits (MAB) that first emerged in Thompson (1933) in the context of drug testing, and was later extended by Robbins (1952) to a more general setting. In this framework, an agent needs to repeatedly choose between $K$ arms, where at each trial the agent pulls one of the arms, and receives a reward. In the formulation of this problem (known as the stochastic MAB setting), rewards from each arm are assumed to be identically distributed and independent across trails and arms. The objective of the agent is to maximize the cumulative return over a certain time horizon, and the performance criterion is the so-called regret: the expected difference between the cumulative reward received by the agent and the reward accumulated by a hypothetical benchmark, referred to as oracle, who holds prior information and the reward distribution of each arm (and thus repeatedly selects the arm with the highest expected reward). Since its inception, this framework has been analyzed under different assumptions to study a variety of applications including clinical trials (Zelen 1969), strategic pricing (Bergemann and Välimäki 1996), packet routing (Awerbuch and Kleinberg 2004), online auctions (Kleinberg and Leighton 2003), online advertising (Pandey et al. 2007), and product recommendations (Madani and DeCoste 2005, Li et al. 2010), among many others.

Classical MAB settings (including the ones used in the above applications) focus on balancing exploration and exploitation in environments where at each period a reward observation is collected only on the arm that is selected by the policy at that time period. However, in many practical settings additional information (that may take various forms, see example below) may become available between decision epochs and may be relevant also to arms that were not selected recently. While in many real-world scenarios utilizing such information flows may be fundamental for achieving good performance, the MAB framework does not account for the extent to which such information flows may impact the design of learning policies and the performance such policies may achieve. We next discuss one concrete application domain to which MAB policies have been commonly applied in the past, and in which such auxiliary information is a fundamental part of the problem.

**The case of cold-start problems in online product recommendations.** Product recommendation systems are widely deployed in the web nowadays, with the objective of helping users navigate through content and consumer products while increasing volume and revenue for service and e-commerce platforms. These systems commonly apply various collaborative filtering and content-based filtering techniques that leverage information such as explicit and implicit preferences of users, product consumption and popularity, and consumer ratings (see, e.g., Hill et al. 1995, Konstan et al. 1997, Breese et al. 1998). While effective when ample information is available on products and consumers, these techniques tend to perform poorly encountering consumers or products that are new to the system and have little or no trace of activity. This phenomenon, termed as the cold-start problem, has been documented...
and studied extensively in the literature; see, e.g., Schein et al. (2002), Park and Chu (2009), and references therein. With this problem in mind, several MAB formulations were suggested and applied for designing recommendation algorithms that effectively balance information acquisition and instantaneous revenue maximization, where arms represent candidate recommendations; see the overview in Madani and DeCoste (2005), as well as later studies by Agarwal et al. (2009), Caron and Bhagat (2013), Tang et al. (2014), and Wang et al. (2017). Aligned with traditional MAB frameworks, these studies consider settings where in each time period observations are obtained only for items that are recommended by the system at that period. However, additional browsing and consumption information may be maintained in parallel to the sequential recommendation process, as a significant fraction of website traffic may take place through means other than recommendation systems (for example, consumers that arrive to product pages directly from external search engines); see browsing data analysis in Sharma and Yan (2013) and Mulpuru (2006), as well as Grau (2009), who estimate that recommendation systems are responsible to only 10-30% of site traffic and revenue. This additional information could potentially be used to better estimate the “reward” from recommending new products and improve the performance of recommendation algorithms facing a cold-start problem (we will discuss in more detail how this could be done after laying out our formulation).

Key challenges and research questions. The availability of additional information (relative to the information collection process that is assumed in classical MAB formulations) fundamentally impacts the design of learning policies and the way a decision maker should balance exploration and exploitation. When additional information is available one may potentially obtain better estimators for the mean rewards, and therefore may need to “sacrifice” less decision epochs for exploration. While this intuition suggests that exploration can be reduced in the presence of additional information, it is a priori not clear how exactly should the appropriate exploration rate depend on the information flows. Moreover, monitoring the exploration levels in real time in the presence of arbitrary information flows introduces additional challenges that have distinct practical relevance. Most importantly, an optimal exploration rate may depend on several characteristics of the information arrival process, such as the amount of information that arrives on each arm, as well as the time at which this information appears (e.g., early on versus later on along the decision horizon). Since it may be hard to predict upfront the salient characteristics of arbitrary information flows, an important challenge is to adapt in real time to an a priori unknown information arrival process and adjust the exploration rate accordingly in order to achieve the performance that is optimal (or near optimal) under prior knowledge of the sample path of information arrivals. This paper is concerned with addressing these challenges.

The main research questions we study in this paper are: (i) How does the best achievable performance (in terms of minimax complexity) that characterizes a sequential decision problem change in the presence

3
of arbitrary information flows? (ii) How should the design of efficient decision-making policies change in the presence of such information flows? (iii) How are achievable performance and policy design affected by the characteristics of the information arrival process, such as the frequency of observations, and their timing? (iv) How can a decision maker adapt to a priori unknown and arbitrary information arrival processes in a manner that guarantees the (near) optimal performance that is achievable under ex-ante knowledge of these processes (“the best of all worlds”)?

1.2 Main contributions

The main contribution of this paper lies in introducing a new, generalized MAB framework with unknown and arbitrary information flows, characterizing the regret complexity of this broad class of MAB problems, and proposing a general policy design approach that demonstrates how effective decision-making policies designed to perform well with very little information can be adjusted in a practical manner that guarantees optimality in information-abundant settings characterized by arbitrary and a priori unknown information flows. More specifically, our contribution is along the following dimensions.

1) **Modeling.** We formulate a new class of MAB problems in the presence of a priori unknown information flows that generalizes the classical stochastic MAB framework, by relaxing strong assumptions that are typically imposed on the information collection process. Our formulation considers a priori unknown information flows that correspond to the different arms and allows information to arrive at arbitrary rate and time. Our formulation therefore captures a large variety of real-world phenomena, yet maintains mathematical tractability.

2) **Analysis.** We establish lower bounds on the performance that is achievable by any non-anticipating policy in the presence of unknown information flows, where performance is measured in terms of regret relative to the performance of an oracle that constantly selects the arm with the highest mean reward. We further show that our lower bounds can be achieved through suitable policy design. These results identify the minimax complexity associated with the MAB problem with unknown information flows, as a function of the information arrivals process, as well as other problem characteristics such as the length of the problem horizon, the number of arms, and parametric characteristics of the family of reward distributions. In particular, we obtain a spectrum of minimax regret rates ranging from the classical regret rates that appear in the stochastic MAB literature when there is no or very little auxiliary information, to a constant regret (independent of the length of the decision horizon) when information arrives frequently and/or early enough.

3) **Policy design.** We introduce a general adaptive exploration approach for designing policies that, without any prior knowledge on the auxiliary information flows, approximate the best performance that is achievable when the information arrival process is known in advance. This “best of all worlds” type of
guarantee implies that rate optimality is achieved uniformly over the general class of information flows at hand (including the case with no information flows, where classical guarantees are recovered).

Our approach relies on using endogenous exploration rates that depend on the amount of information that becomes available over time. In particular, it is based on adjusting in real time the effective exploration rate of MAB policies that were designed to perform well in the absence of any auxiliary information flows, while leveraging the structure of said policies. More precisely, various well-known MAB policies govern the rate at which sub-optimal options are explored through some monotonously decreasing function of the time period, where the precise structure of said function may change from one algorithm to another. Our approach leverages the optimality of these functional structures in the absence of any auxiliary information flows, while replacing the time index with virtual time indexes that are dynamically updated based on information arrivals. Whenever auxiliary information on a certain arm arrives, the virtual time index that is associated with that arm is advanced using a carefully selected multiplicative factor, and thus the rate at which the policy is experimenting with that arm is reduced. We demonstrate the effectiveness and practicality of the adaptive exploration approach through establishing performance bounds and evaluating numerically the performance of the adjusted versions of well-known MAB policies.

(4) Reactive information flows. Our formulation focuses on information flows that are arbitrary and unknown to the decision maker, but are fixed upfront and independent of the decision path of the policy. In §6 of the paper we extend our framework to consider a broad class of information flows that are reactive to the past actions of the decision-making policy. We study the impact endogenous information flows may have on the achievable performance, and establish the optimality of the adaptive exploration approach for a broad class of endogenous information flows.

1.3 Related work

Multi-armed bandits. For a comprehensive overview of MAB formulations we refer the readers to the monographs by Berry and Fristedt (1985) and Gittins et al. (2011) for Bayesian / dynamic programming formulations, as well as to Cesa-Bianchi and Lugosi (2006) and Bubeck et al. (2012) that cover the machine learning literature and the so-called adversarial setting. A sharp regret characterization for the more traditional framework (random rewards realized from stationary distributions), often referred to as the stochastic MAB problem, was first established by Lai and Robbins (1985), followed by analysis of important policies designed for the stochastic framework, such as the $\epsilon$-greedy, UCB1, and Thompson sampling; see, e.g., Auer, Cesa-Bianchi, and Fischer (2002), as well as Agrawal and Goyal (2013).

The MAB framework focuses on balancing exploration and exploitation, typically under very little assumptions on the distribution of rewards, but with very specific assumptions on the future information
collection process. In particular, optimal policy design is typically predicated on the assumption that at each period a reward observation is collected only on the arm that is selected by the policy at that time period (exceptions to this common information structure are discussed below). In that sense, such policy design does not account for information (e.g., that may arrive between pulls) that may be available in many practical settings, and that might be essential for achieving good performance. In the current paper we relax the information structure of the classical MAB framework by allowing arbitrary information arrival processes. Our focus is on: (i) studying the impact of the information arrival characteristics (such as frequency and timing) on policy design and achievable performance; and (ii) adapting to a priori unknown sample path of information arrivals in real time.

As alluded to above, there are few MAB settings (and other sequential decision frameworks) in which more information can be collected in each time period. One example is the so-called contextual MAB setting, also referred to as bandit problem with side observations (Wang et al. 2005), or associative bandit problem (Strehl et al. 2006), where at each trial the decision maker observes a context carrying information about other arms. Another important example is the full-information adversarial MAB setting, where rewards are not characterized by a stationary stochastic process but are rather arbitrary and can even be selected by an adversary (Auer et al. 1995, Freund and Schapire 1997). In the full-information adversarial MAB setting, at each time period the agent not only observes the reward generated by the arm that was selected, but also observes the rewards generated by rest of the arms. While the adversarial nature of the latter setting makes it fundamentally different in terms of achievable performance, analysis, and policy design, from the stochastic formulation that is adopted in this paper, it is also notable that the above settings consider very specific information structures that are a priori known to the agent, as opposed to our formulation where the characteristics of the information flows are arbitrary and a priori unknown.

**Balancing and regulating exploration.** Several papers have considered different settings of sequential optimization with partial information and distinguished between cases where exploration is unnecessary (a myopic decision-making policy achieves optimal performance), and cases where exploration is essential for achieving good performance (myopic policies may lead to incomplete learning and large losses); see, e.g., Harrison et al. (2012) and den Boer and Zwart (2013) that study policies for dynamic pricing without knowing the demand function, Besbes and Muharremoglu (2013) for inventory management without knowing the demand distribution, and Lee et al. (2003) in the context of technology development. In a recent paper, Bastani et al. (2017) consider the contextual MAB framework and show that if the distribution of the contextual information guarantees sufficient diversity, then exploration becomes unnecessary and greedy policies can benefit from the natural exploration that is embedded in the information diversity to achieve asymptotic optimality. In related studies, Woodroofe (1979) and Sarkar
consider a Bayesian one armed contextual MAB problem and show that a myopic policy achieves asymptotic optimal when the discount factor converges to one.

On the other hand, few papers have studied cases where exploration is not only essential but should be particularly frequent in order to maintain optimality. For example, Besbes et al. (2014) consider a general MAB framework where the reward distribution may change over time according to a budget of variation, and characterize the manner in which optimal exploration rates increase as a function of said budget. In addition, Shah et al. (2018) consider a platform in which the preferences of arriving users may depend on the experience of previous users. They show that in such setting classical MAB policies may under-explore, and introduce a balanced-exploration approach that results in optimal performance.

The above studies demonstrate a variety of practical settings where the extent of exploration that is required to maintain optimality strongly depends on particular problem characteristics that may often be a priori unknown to the decision maker. This introduces the challenge of endogenizing exploration: dynamically adapting the rate at which a decision-making policy explores to identify the appropriate rate of exploration and to approximate the best performance that is achievable under ex ante knowledge of the underlying problem characteristics. In this paper we address this challenge from information collection perspective. We identify conditions on the information arrival process that guarantee the optimality of myopic policies, and further identify adaptive MAB policies that guarantee (near) optimal performance without prior knowledge on the information arrival process (“best of all worlds”).

In addition, few papers considered approaches of regulating exploration rates based on a priori known characteristics of payoff structure, in settings that are different than ours. For example, Tracà and Rudin (2015) consider an approach of regulating exploration in a setting where rewards are scaled by an exogenous multiplier that temporally evolves in an a priori known manner, and show that in such setting the performance of known MAB policies can be improved if exploration is increased in periods of low reward. Another approach of regulating the exploration is studied by Komiyama et al. (2013) in a setting that includes lock-up periods in which the agent cannot change her actions.

Adaptive algorithms. One of the challenges we address in this paper lies in designing policies that adapt in real time to the arrival process of information, in the sense of achieving ex-post performance which is as good (or nearly as good) as the one achievable under ex-ante knowledge on the information arrival process. This challenge dates back to studies in the statistics literature (see Tsybakov (2008) and references therein), and has seen recent interest in the machine learning and sequential decision making literature streams; examples include Seldin and Slivkins (2014) that present an algorithm that achieves (near) optimal performance in both stochastic and adversarial multi-armed bandit regimes without prior knowledge on the nature of environment, Sani et al. (2014) that consider an online convex optimization setting and derive algorithms that are rate optimal regardless of whether the target function is weakly
or strongly convex, Jadabaie et al. (2015) that study the design of adaptive algorithms that compete against dynamic benchmarks, and Luo and Schapire (2015) that address the problem of learning under experts’ advices and compete in an adversarial setting against any convex combination of experts.

2 Problem formulation

In this section we formulate a class of multi-armed bandit problems with auxiliary information flows. We note that many of our modeling assumptions can be generalized and are made only to simplify exposition and analysis. Some generalizations of our formulation are discussed in §.

Let $K = \{1, \ldots, K\}$ be a set of arms (actions) and let $T = \{1, \ldots, T\}$ denote a sequence of decision epochs. At each time period $t \in T$, a decision maker selects one of the $K$ arms. When selecting an arm $k \in K$ at time $t \in T$, a reward $X_{k,t} \in \mathbb{R}$ is realized and observed. For each $t \in T$ and $k \in K$, the reward $X_{k,t}$ is assumed to be independently drawn from some $\sigma^2$-sub-Gaussian distribution with mean $\mu_k$.\footnote{A real-valued random variable $X$ is said to be sub-Gaussian if there is some $\sigma > 0$ such that for every $\lambda \in \mathbb{R}$ one has $\mathbb{E}e^{\lambda(X - \mathbb{E}X)} \leq e^{\sigma^2 \lambda^2 / 2}$.}

We denote the profile of rewards at time $t$ by $X_t = (X_{1,t}, \ldots, X_{K,t})^\top$ and the profile of mean-rewards by $\mu = (\mu_1, \ldots, \mu_K)^\top$. We further denote by $\nu = (\nu_1, \ldots, \nu_K)^\top$ the distribution of the rewards profile $X_t$. We assume that rewards are independent across time periods and arms. We denote the highest expected reward and the best arm by $\mu^*$ and $k^*$ respectively, that is\footnote{This broad class of distributions include, for instance, Gaussian random variables, as well as any random variable with a bounded support (if $X \in [a, b]$ then $X$ is $\frac{|a-b|}{4}$-sub-Gaussian) such as Bernoulli random variables. Notably, if a random variable is $\sigma^2$-sub-Gaussian, it is also $\tilde{\sigma}^2$-sub-Gaussian for all $\tilde{\sigma} > \sigma$.}

$$\mu^* = \max_{k \in K}\{\mu_k\}, \quad k^* = \arg\max_{k \in K} \mu_k.$$\footnote{For the sake of simplicity, in the formulation and hereafter in the rest of the paper when using the arg min and arg max operators we assume that ties are broken in favor of the smaller index.}

We denote by $\Delta_k = \mu^* - \mu_k$ the difference between the expected reward of the best arm and the expected reward of arm $k$. We assume prior knowledge of a positive lower bound $0 < \Delta \leq \min_{k \in K \setminus \{k^*\}} \Delta_k$ as well as a positive number $\sigma > 0$ for which all the reward distributions are $\sigma^2$-sub-Gaussian. We denote by $S = S(\Delta, \sigma^2)$ the class of $\Delta$-separated $\sigma^2$-sub-Gaussian distribution profiles:

$$S(\Delta, \sigma^2) := \left\{ \nu = (\nu_1, \ldots, \nu_K)^\top \mid \Delta \cdot \mathbb{1}\{k \neq k^*\} \leq \Delta_k \text{ and } \mathbb{E}e^{\lambda(X_{k,1} - \mu_k)} \leq e^{\sigma^2 \lambda^2 / 2} \ \forall k \in K, \forall \lambda \in \mathbb{R} \right\}.$$\footnote{For the sake of simplicity, in the formulation and hereafter in the rest of the paper when using the arg min and arg max operators we assume that ties are broken in favor of the smaller index.}

Auxiliary information flows. Before each round $t$, the agent may or may not observe reward realizations for some of the arms without pulling them. Let $\eta_{k,t} \in \{0, 1\}$ denote the indicator of observing an auxiliary information on arm $k$ just before time $t$. We denote by $\eta_t = (\eta_{1,t}, \ldots, \eta_{K,t})^\top$
the vector of indicators $\eta_{k,t}$’s associated with time step $t$, and by $H = (\eta_1, \ldots, \eta_T)$ the information arrival matrix with columns $\eta_t$’s; we assume that this matrix is independent of the policy’s actions (this assumption will be relaxed later on). If $\eta_{k,t} = 1$, then a random variable $Y_{k,t} \sim \nu_k$ is observed. We denote $Y_t = (Y_{1,t}, \ldots, Y_{K,t})^\top$, and assume that the random variables $Y_{k,t}$ are independent across time periods and arms and are also independent from the reward realizations $X_{k,t}$. We denote the vector of new information received just before time $t$ by $Z_t = (Z_{1,t}, \ldots, Z_{K,t})^\top$ where for any $k$ one has:

$$Z_{k,t} = \eta_{k,t} \cdot Y_{k,t}.$$

**Admissible policies, performance, and regret.** Let $U$ be a random variable defined over a probability space $(\mathbb{U}, \mathcal{U}, P_u)$. Let $\pi_t : \mathbb{R}^{t-1} \times \mathbb{R}^{K \times t} \times \{0, 1\}^{K \times t} \times \mathbb{U} \to \mathcal{K}$ for $t = 1, 2, 3, \ldots$ be measurable functions (with some abuse of notation we also denote the action at time $t$ by $\pi_t \in \mathcal{K}$) given by

$$\pi_t = \begin{cases} \pi_1(Z_1, \eta_1, U) & t = 1, \\ \pi_t(X_{\pi_{t-1}, t-1}, \ldots, X_{\pi_1, 1}, Z_t, \ldots, Z_1, \eta_t, \ldots, \eta_1, U) & t = 2, 3, \ldots \end{cases}$$

The mappings $\{\pi_t : t = 1, \ldots, T\}$, together with the distribution $P_u$ define the class of admissible policies. We denote this class by $P$. We further denote by $\{\mathcal{H}_t, t = 1, \ldots, T\}$ the filtration associated with a policy $\pi \in P$, such that $\mathcal{H}_1 = \sigma(Z_1, \eta_1, U)$, and $\mathcal{H}_t = \sigma(\{X_{\pi_s, s}\}_{s=1}^{t-1}, \{Z_s\}_{s=1}^t, \{\eta_s\}_{s=1}^t, \mathbb{U})$ for all $t \in \{2, 3, \ldots\}$. Note that policies in $P$ depend only on the past history of actions and observations as well as auxiliary information arrivals, and allow for randomization via their dependence on $U$.

We evaluate the performance of a policy $\pi \in P$ by the regret it incurs under information arrival process $H$ relative to the performance of an oracle that selects the arm with the highest expected reward. We define the worst-case regret as follows:

$$R^*_S(H, T) = \sup_{\nu \in S} E^\pi_{\nu} \left[ \sum_{t=1}^T (\mu^* - \mu^{\pi_t}) \right],$$

where the expectation $E^\pi_{\nu} [\cdot]$ is taken with respect to the noisy rewards, as well as to the policy’s actions (throughout the paper we will denote by $P^\pi_{\nu}$, $E^\pi_{\nu}$, and $R^*_S$ the probability, expectation, and regret when the arms are selected according to policy $\pi$ and rewards are distributed according to $\nu$). In addition, we denote by $R^*_S(H, T) = \inf_{\pi \in P} R^*_S(H, T)$ the best achievable guaranteed performance: the minimal regret that can be guaranteed by an admissible policy $\pi \in P$. In the following sections we study the magnitude of $R^*_S(H, T)$ as a function of the information arrival process $H$. 


2.1 Discussion of model assumptions and extensions

2.1.1 Generalized information structure and the cold-start problem

In our formulation we focus on auxiliary observations that have the same distribution as reward observations, but all our results hold for a broad family of information structures as long as unbiased estimators of mean rewards can be constructed from the auxiliary observations, that is, when there exists a mapping $\phi(\cdot)$ such that $\mathbb{E}[\phi(Y_{k,t})] = \mu_k$ for each $k$. Notably, the compatibility with such generalized information structure allows a variety of concrete practical applications of our analysis, including the cold-start problem referred to in §1.1.

In particular, in the context of a product recommendation system, the reward of recommending an item $k$ is typically proportional to both the click-through rate (CTR) of item $k$, measuring the tendency of users to click on the recommendations, as well as the conversion rate (CVR) of that item, measuring the payoff from the actions taken by the user after the click (for example, average revenue spent once a user arrives to the product page); namely, it is common to assume that $\mu_k = \mathbb{E}X_{k,t} = c_k \cdot \text{CTR}_k \cdot \text{CVR}_k$, for some constant $c_k$. Notably, several methods have been developed to estimate click-through rates in the absence of historical information on new products. These include MAB-based methods that were discussed in §1.1, as well as other methods such as ones that focus on feature-based or semantic-based decomposition of products; see, e.g., Richardson et al. (2007) who focus on feature-based analysis, as well as Regelson and Fain (2006) and Dave and Varma (2010) that estimate click-through rates based on semantic context.

However, since click-through rates are typically low (and since clicks are prerequisites for conversions), and due to a common delay between clicks and conversions, estimating conversion rates may be still challenging long after reasonable click-through estimates were already established; see, e.g., Chapelle (2014), Lee et al. (2013), Lee et al. (2012), and Rosales et al. (2012), as well as the empirical analysis and comparison of several recommendations methods in Zhang et al. (2014). Leveraging this time-scale separation, and given estimated click-through probabilities $\text{CTR}_1, \ldots, \text{CTR}_K$, our formulation allows system optimization through online learning of the mean rewards $\{\mu_k\}$ not only from recommending the different items and observing rewards $X_{k,t}$ (such that $\mu_k = \mathbb{E}X_{k,t}$) of recommended products, but also through auxiliary unbiased estimators of the form $\text{CTR}_k \cdot Y_{k,t}$ (such that $\mu_k = \text{CTR}_k \cdot \mathbb{E}Y_{k,t}$), where $Y_{k,t}$ is the revenue spent by a user arriving to a product page of item $k$ through means other than the recommendation system itself (e.g., from an external search engine).

2.1.2 Other generalizations and extensions

For the sake of simplicity, our model adopts the basic and well studied stochastic MAB framework (Lai and Robbins 1985). However, our methods and analysis can be directly applied to more general
frameworks such as the contextual MAB where mean rewards are linearly dependent on context vectors; see, e.g., Goldenshluger and Zeevi (2013) and references therein.

For the sake of simplicity we assume that only one information arrival can occur before each time step for each arm (that is, for each time $t$ and arm $k$, one has that $\eta_{k,t} \in \{0,1\}$). Notably, all our results can be extended to allow more than one information arrival per time step per arm.

We focus on a setting where the information arrival process (namely, the matrix $H$) is unknown, yet fixed and independent of the sequence of decisions and observations. While fully characterizing the regret complexity when information flows may depend on the history is a challenging open problem, in §6 we analyze the optimal exploration rate, and study policy design under a broad class of information flows that are reactive to the past decisions of the policy.

We finally note that for the sake of simplicity we refer to the lower bound $\Delta$ on the differences in mean rewards relative to the best arm as a fixed parameter that is independent of the horizon length $T$. This corresponds to the case of separable mean rewards, which is prominent in the classical stochastic MAB literature. Nevertheless, we do not make any explicit assumption on the separability of mean rewards and note that our analysis and results hold for the more general case where the lower bound $\Delta$ is a function of the horizon length $T$. This includes the case where mean rewards are not separable, in the sense that $\Delta$ is decreasing with $T$.

### 3 The impact of information flows on achievable performance

In this section we study the impact auxiliary information flows may have on the performance that one could aspire to achieve. Our first result formalizes what cannot be achieved, establishing a lower bound on the best achievable performance as a function of the information arrival process.

**Theorem 1 (Lower bound on the best achievable performance)** For any $T \geq 1$ and information arrival matrix $H$, the worst-case regret for any admissible policy $\pi \in \mathcal{P}$ is bounded below as follows

$$R_\mathcal{S}(H,T) \geq \frac{C_1}{\Delta} \sum_{k=1}^K \left( \frac{C_2 \Delta^2}{K} \sum_{t=1}^T \exp \left( -C_3 \Delta^2 \sum_{s=1}^t \eta_{s,k} \right) \right),$$

where $C_1$, $C_2$, and $C_3$ are positive constants that only depend on $\sigma$.

The precise expressions of $C_1$, $C_2$, and $C_3$ are provided in the discussion below. Theorem 1 establishes a lower bound on the achievable performance in the presence of unknown information flows. This lower bound depends on an arbitrary sample path of information arrivals, captured by the elements of the matrix $H$. In that sense, Theorem 1 provides a spectrum of bounds on achievable performances, mapping many potential information arrival trajectories to the best performance they allow. In particular, when
there is no additional information over what is assumed in the classical MAB setting (that is, when $H = 0$), we recover a lower bound of order $\frac{K}{\Delta} \log T$ that coincides with the bounds established in Lai and Robbins (1985) and Bubeck et al. (2013) for that setting. Theorem 1 further establishes that when additional information is available, achievable regret rates may become lower, and that the impact of information arrivals on the achievable performance depends on the frequency of these arrivals, but also on the time at which these arrivals occur; we further discuss these observations in §3.1.

**Key ideas in the proof.** The proof of Theorem 1 adapts to our framework ideas of identifying a worst-case nature “strategy”; see, e.g., the proof of Theorem 6 in Bubeck et al. (2013). While the full proof is deferred to the appendix, we next illustrate its key ideas using the special case of two arms. We consider two possible profiles of reward distributions, $\nu$ and $\nu'$, that are “close” enough in the sense that it is hard to distinguish between the two, but “separated” enough such that a considerable regret may be incurred when the “correct” profile of distributions is misidentified. In particular, we assume that the decision maker is a priori informed that the first arm generates rewards according to a normal distribution with standard variation $\sigma$ and a mean that is either $-\Delta$ (according to $\nu$) or $+\Delta$ (according to $\nu'$), and the second arm is known to generate rewards with normal distribution of standard variation $\sigma$ and mean zero. To quantify a notion of distance between the possible profiles of reward distributions we use the Kullback-Leibler (KL) divergence. The KL divergence between two positive measures $\rho$ and $\rho'$ with $\rho$ absolutely continuous with respect to $\rho'$, is defined as:

$$
\text{KL}(\rho, \rho') := \int \log \left( \frac{d\rho}{d\rho'} \right) d\nu = \mathbb{E}_\rho \log \left( \frac{d\rho}{d\rho'} (X) \right),
$$

where $\mathbb{E}_\rho$ denotes the expectation with respect to probability measure $\rho$. Using Lemma 2.6 from Tsybakov (2008) that connects the KL divergence to error probabilities, we establish that at each period $t$ the probability of selecting a suboptimal arm must be at least

$$
p_{t,\text{sub}} = \frac{1}{4} \exp \left( -\frac{2\Delta^2}{\sigma^2} \left( \mathbb{E}_\nu[\tilde{n}_{1,t}] + \sum_{s=1}^{t} \eta_{1,s} \right) \right),
$$

where $\tilde{n}_{1,t}$ denotes the number of times the first arm is pulled up to time $t$. Each selection of suboptimal arm contributes $\Delta$ to the regret, and therefore the cumulative regret must be at least $\Delta \sum_{t=1}^{T} p_{t,\text{sub}}$. We further observe that if arm 1 has mean rewards of $-\Delta$, the cumulative regret must also be at least $\Delta \cdot \mathbb{E}_\nu[\tilde{n}_{1,T}]$. Therefore the regret is lower bounded by $\frac{\Delta}{2} \left( \sum_{t=1}^{T} p_{t,\text{sub}} + \mathbb{E}_\nu[\tilde{n}_{1,T}] \right)$ which is greater than $\frac{\sigma^2}{2\Delta} \log \left( \frac{\Delta^2}{2\Delta} \sum_{t=1}^{T} \exp \left( -\frac{2\Delta^2}{\sigma^2} \sum_{s=1}^{t} \eta_{1,s} \right) \right)$. The argument can be repeated by switching arms 1 and 2. For
\[ K \text{ arms, we follow the above lines and average over the established bounds to obtain:} \]
\[
\mathcal{R}_S^\pi(H, T) \geq \frac{\sigma^2(K - 1)}{4K \Delta} \sum_{k=1}^K \log \left( \frac{\Delta^2}{\sigma^2K} \sum_{t=1}^T \exp \left( -\frac{2\Delta^2}{\sigma^2} \sum_{s=1}^t \eta_{s,k} \right) \right),
\]
which establishes the result. ■

3.1 Discussion and subclasses of information flows

Theorem 1 demonstrates that information flows may be leveraged to improve performance and reduce regret rates, and that their impact on the achievable performance increases when information arrives more frequently, and earlier. This observation is consistent with the following intuition: (i) at early time periods we have collected only few observations and therefore the marginal impact of an additional observation on the stochastic error rates is large; and (ii) when information appears early on, there are more future opportunities where this information can be used. To emphasize this observation we next demonstrate the implications on achievable performance of two concrete information arrival processes of natural interest: a process with a fixed arrival rate, and a process with a decreasing arrival rate.

3.1.1 Stationary information flows

Assume that \( \eta_{k,t} \)'s are i.i.d. Bernoulli random variables with mean \( \lambda \). Then, for any \( T \geq 1 \) and admissible policy \( \pi \in \mathcal{P} \), one obtains the following lower bound for the achievable performance:

1. If \( \lambda \leq \frac{\sigma^2}{4\Delta^2T} \), then
   \[
   \mathbb{E}_H \left[ \mathcal{R}_S^\pi(H, T) \right] \geq \frac{\sigma^2(K - 1)}{4\Delta} \log \left( \frac{1 - e^{-1/2}}{\sigma^2K} \frac{\Delta^2T}{\lambda K} \right).\n   \]

2. If \( \lambda \geq \frac{\sigma^2}{4\Delta^2T} \), then
   \[
   \mathbb{E}_H \left[ \mathcal{R}_S^\pi(H, T) \right] \geq \frac{\sigma^2(K - 1)}{4\Delta} \log \left( \frac{1 - e^{-1/2}}{2\lambda K} \right).\n   \]

This class includes instances in which, on average, information arrives at a constant rate \( \lambda \). Analyzing these arrival process reveals two different regimes. When the information arrival rate is small enough, auxiliary observations become essentially ineffective, and one recovers the performance bounds that were established for the classical stochastic MAB problem. In particular, as long as there are no more than order \( \Delta^{-2} \) information arrivals over \( T \) time periods, this information does not impact achievable regret rates.\(^3\) When \( \Delta \) is fixed and independent of the horizon length \( T \), the lower bound scales logarithmically

\(^3\)This coincides with the observation that one requires order \( \Delta^{-2} \) samples to distinguish between two distributions that are \( \Delta \)-separated; see, e.g., [Audibert and Bubeck 2010].
with \( T \). When \( \Delta \) can scale with \( T \), a bound of order \( \sqrt{T} \) is recovered when \( \Delta \) is of order \( T^{-1/2} \). In both cases, there are known policies (such as UCB1) that guarantee rate-optimal performance; for more details see policies, analysis, and discussion in [Auer et al. (2002)](https://www.jstor.org/stable/30036056).

On the other hand, when there are more than order \( \Delta^{-2} \) observations over \( T \) periods, the lower bound on the regret becomes a function of the arrival rate \( \lambda \). When the arrival rate is independent of the horizon length \( T \), the regret is bounded by a constant that is independent of \( T \), and a myopic policy (e.g., a policy that for the first \( K \) periods pulls each arm once, and at each later period pulls the arm with the current highest estimated mean reward, while randomizing to break ties) is optimal. For more details see sections C.1 and C.2 of the Appendix.

### 3.1.2 Diminishing information flows

Fix some \( \kappa > 0 \), and assume that \( \eta_{k,t} \)'s are random variables such that for each arm \( k \in \mathcal{K} \) and at each time step \( t \),

\[
\mathbb{E} \left[ \sum_{s=1}^{t} \eta_{k,s} \right] = \frac{\sigma^2 \kappa}{2\Delta^2} \log t.
\]

Then, for any \( T \geq 1 \) and admissible policy \( \pi \in \mathcal{P} \), one obtains the following lower bound for the achievable performance:

1. If \( \kappa < 1 \) then:
   \[
   \mathcal{R}_S^\pi(H, T) \geq \frac{\sigma^2(K - 1)}{4\Delta} \log \left( \frac{\Delta^2 / K \sigma^2}{1 - \kappa} (T + 1)^{1-\kappa} - 1 \right).
   \]

2. If \( \kappa > 1 \) then:
   \[
   \mathcal{R}_S^\pi(H, T) \geq \frac{\sigma^2(K - 1)}{4\Delta} \log \left( \frac{\Delta^2 / K \sigma^2}{\kappa - 1} \left( 1 - \frac{1}{(T + 1)^{\kappa - 1}} \right) \right).
   \]

This class includes information flows under which the expected number of information arrivals up to time \( t \) is of order \( \log t \). This class demonstrates the impact of the timing of information arrivals on the achievable performance, and suggests that a constant regret may be achieved even when the rate of information arrivals is decreasing. Whenever \( \kappa < 1 \), the lower bound on the regret is logarithmic in \( T \), and there are well-studied MAB policies (e.g., UCB1, [Auer et al. (2002)](https://www.jstor.org/stable/30036056)) that guarantee rate-optimal performance. When \( \kappa > 1 \), the lower bound on the regret is a constant, and one may observe that when \( \kappa \) is large enough a myopic policy is asymptotically optimal. (In the limit \( \kappa \to 1 \) the lower bound is of order \( \log \log T \).) For more details see sections C.3 and C.4 of the Appendix.

### 3.1.3 Discussion

One may contrast the classes of information flows described in §3.1.1 and §3.1.2 by selecting \( \kappa = \frac{2\Delta^2 \lambda T}{\sigma^2 \log T} \). Then, in both settings the total number of information arrivals for each arm is \( \lambda T \). However, while in
the first class the information arrival rate is fixed over the horizon, in the second class this arrival rate is higher in the beginning of the horizon and gradually decreasing over time. The different timing of the $\lambda T$ information arrivals may lead to different regret rates. To demonstrate this, further select $\lambda = \frac{\sigma^2 \log T}{\Delta^2}$, which implies $\kappa = 2$. The lower bound in §3.1.1 is then logarithmic in $T$ (establishing the impossibility of constant regret in that setting), but the lower bound in §3.1.2 is constant and independent of $T$ (in the next section we will see that constant regret is indeed achievable in this setting). This observation echoes the intuition that earlier observations have higher impact on achievable performance, as at early periods there is only little information that is available and therefore the marginal impact of an additional observation on the performance is larger, and since earlier information can be used for more decision periods (as the remaining horizon is longer).

The analysis above demonstrates that optimal policy design and the best achievable performance depend on the information arrival process: while policies such as UCB1 and $\epsilon$-greedy, that explore over arms (and in that sense are not myopic) may be rate optimal in some cases, a myopic policy that does not explore (except perhaps in a small number of periods in the beginning of the horizon) can achieve rate-optimal performance in other cases. However, the identification of a rate-optimal policy relies on prior knowledge of the information flow. Therefore, an important question one may ask is: How can a decision maker adapt to an arbitrary and unknown information arrival process in the sense of achieving (near) optimal performance without any prior knowledge on the information flow? We address this question in the following sections.

4 General approach for designing near-optimal adaptive policies

In this section we suggest a general approach for adapting to a priori unknown information flows. Before laying down our approach, we first demonstrate that classical policy design may fail to achieve the lower bound in Theorem 1 in the presence of unknown information flows.

The inefficiency of naive adaptations of MAB policies. Consider a simple approach of adapting classical MAB policies to account for arriving information when calculating the estimates of mean rewards, while maintaining the structure of the policy otherwise. Such an approach can be implemented

$$\mathbb{E} \left[ \sum_{s=1}^{t} \eta_{k,s} \right] = \lambda T \frac{t^{1-\gamma} - 1}{T^{1-\gamma} - 1}.$$  

The expected number of total information arrivals for each arm, $\lambda T$, is determined by the parameter $\lambda$. The concentration of arrivals, however, is governed by the parameter $\gamma$. When $\gamma = 0$ the arrival rate is constant, corresponding to the class described in §3.1.1. As $\gamma$ increases, information arrivals concentrate in the beginning of the horizon, and $\gamma \to 1$ leads to $\mathbb{E} \left[ \sum_{s=1}^{t} \eta_{k,s} \right] = \lambda T \frac{\log t}{\log T}$, which corresponds to the class in §3.1.2. Then, one may apply similar analysis to observe that when $\lambda T$ is of order $T^{1-\gamma}$ or higher, the lower bound is a constant independent of $T$. 

using well-known MAB policies such as UCB1 or epsilon-greedy. One observation is that the performance bounds of these policies (analyzed, e.g., in Auer et al. 2002) do not improve (as a function of the horizon length $T$) in the presence of unknown information flows. Moreover, it is possible to show through lower bounds on the guaranteed performance that these policies indeed achieve sub-optimal performance. To demonstrate this, consider the subclass of stationary information flows described in §3.1.1 with an arrival rate $\lambda$ that is large compared to $\frac{\sigma^2}{4\Delta^2 T}$. In that case, we have seen that the regret lower bound becomes constant whenever the arrival rate $\lambda$ is independent of $T$. However, the $\epsilon$-greedy policy, employs an exploration rate that is independent of the number of observations that were obtained for each arm and therefore effectively incurs regret of order $\log T$ due to performing unnecessary exploration.

**A simple rate-optimal policy.** To advance our approach we provide a simple and deterministic adaptive exploration policy that includes the key elements that are essential for appropriately adjusting the exploration rate and achieving good performance in the presence of unknown information flows. In what follows, we denote by $n_{k,t}$ and $X_{k,n_{k,t}}$ the number of times a sample from arm $k$ has been observed and the empirical average reward of arm $k$ up to time $t$, respectively, that is,

$$n_{k,t} = \eta_t + \sum_{s=1}^{t-1} (\eta_{k,s} + 1\{\pi_s = k\}), \quad X_{k,n_{k,t}} = \frac{\eta_{k,t} Y_{k,t} + \sum_{s=1}^{t-1} (\eta_{k,s} Y_{k,s} + 1\{\pi_s = k\} X_{k,s})}{n_{k,t}}.$$

Consider the following policy:

<table>
<thead>
<tr>
<th>Adaptive exploration policy. Input: a tuning parameter $c &gt; 0$.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Initialization. Set initial virtual times $\tau_{k,0} = 0$ for all $k \in \mathcal{K}$, and an exploration set $\mathcal{W}_0 = \mathcal{K}$.</td>
</tr>
<tr>
<td>2. At each period $t = 1, 2, \ldots, T$:</td>
</tr>
<tr>
<td>(a) Observe the vectors $\eta_t$ and $Z_t$.</td>
</tr>
<tr>
<td>- Advance virtual times indexes: $\tau_{k,t} = (\tau_{k,t-1} + 1) \cdot \exp \left( \frac{\eta_{k,t} \Delta^2}{\epsilon \sigma^2} \right)$ for all $k \in \mathcal{K}$.</td>
</tr>
<tr>
<td>- Update the exploration set: $\mathcal{W}<em>t = \left{ k \in \mathcal{K} \mid n</em>{k,t} &lt; \frac{\epsilon \sigma^2 \Delta^2}{\Delta^2} \log \tau_{k,t} \right}$.</td>
</tr>
<tr>
<td>(b) If $\mathcal{W}_t$ is not empty, select an arm from $\mathcal{W}<em>t$ with the fewest observations: (exploration) $\pi_t = \arg \min</em>{k \in \mathcal{W}<em>t} n</em>{k,t}$.</td>
</tr>
<tr>
<td>Otherwise, Select an arm with the highest estimated reward: (exploitation) $\pi_t = \arg \max_{k \in \mathcal{K}} X_{k,n_{k,t}}$.</td>
</tr>
<tr>
<td>In both cases, let ties be broken in favor of the arm with the lowest $k$ index.</td>
</tr>
</tbody>
</table>
(c) Receive and observe a reward $X_{\pi_t,t}$

Clearly $\pi \in \mathcal{P}$. At each time step, the adaptive exploration policy checks for each arm $k$ whether the number of observations that has been collected so far (through arm pulls and auxiliary information together) exceeds a dynamic threshold that depends logarithmically on a virtual time index $\tau_{k,t}$, that is, whether arm $k$ satisfies the condition $n_{k,t} \geq \frac{c\sigma^2}{\Delta^2} \log \tau_{k,t}$. If yes, the arm with the highest reward estimator $\bar{X}_{k,n_{k,t}}$ is pulled (exploitation). Otherwise, the arm with the fewest observations is pulled (exploration). The condition $n_{k,t} \geq \frac{c\sigma^2}{\Delta^2} \log \tau_{k,t}$, when satisfied by all arms, guarantees that enough observations have been collected from each arm such that a suboptimal arm will be selected with a probability of order $t^{-c/8}$ or less (a rigorous derivation appears in the proof of Theorem 2).

The adaptive exploration policy generalizes a principle of balancing exploration and exploitation that is common in the absence of auxiliary information flows, by which the exploration rate is set in a manner that guarantees that the overall loss due to exploration would equal the expected loss due to misidentification of the best arm; see e.g., Auer et al. (2002) and references therein, the related concept of forced sampling in Langford and Zhang (2008), as well as related discussions in Goldenshluger and Zeevi (2013) and Bastani and Bayati (2015). In the absence of auxiliary information flows, a decreasing exploration rate of order $1/t$ guarantees that the arm with the highest estimated mean reward can be suboptimal only with a probability of order $1/t$; see, e.g., the analysis of the $\epsilon$-greedy policy in Auer et al. (2002), where at each time period $t$ exploration occurs uniformly at random with probability $1/t$.

Recalling the discussion in §3.1.3, the decay of exploration rates over time captures the manner in which new information becomes less valuable over time.

In the presence of additional information stochastic error rates may decrease. The adaptive exploration policy dynamically reacts to the information flows by effectively reducing the exploration rates for different arms to guarantee that the loss due to exploration is balanced throughout the horizon with the expected loss due to misidentification of the best arm. This balance is kept by adjusting virtual time indexes $\tau_{k,t}$ that are associated with each arm (replacing the actual time index $t$, which is appropriate in the absence of auxiliary information flows). In particular, the adaptive exploration policy explores each arm $k$ at a rate that would have been appropriate without auxiliary information flows at a future time step $\tau_{k,t}$. Every time additional information on arm $k$ is observed, a carefully selected multiplicative factor is used to further advance the virtual time index $\tau_{k,t}$ according to the update rule:

$$\tau_{k,t} = (\tau_{k,t-1} + 1) \cdot \exp(\delta \cdot \eta_{k,t}),$$

for some suitably selected $\delta$. The general idea of adapting the exploration rate of a policy by advancing
Figure 1: Illustration of the adaptive exploration approach. (Left) Virtual time index $\tau$ is advanced using multiplicative factors whenever auxiliary information is observed. (Right) Exploration rate decreases as a function of $\tau$, and in particular, exhibits discrete “jumps” whenever auxiliary information is observed.

The following result characterizes the performance that is guaranteed by the adaptive exploration policy in the presence of auxiliary information flows.

**Theorem 2 (Near optimality of adaptive exploration policy)** Let $\pi$ be an adaptive exploration policy tuned by $c > 8$. For any $T \geq 1$ and information arrival matrix $H$:

$$R_S^\pi(H, T) \leq \sum_{k \in \mathcal{K}} \Delta_k \left( \frac{C_4}{\Delta^2} \log \left( \sum_{t=1}^T \exp \left( -\frac{\Delta^2}{C_4} \sum_{s=1}^t \eta_{k,s} \right) \right) + C_5 \right),$$

where $C_4$ and $C_5$ are positive constants that depend only on $\sigma$.

**Key ideas in the proof.** To establish the result we decompose the regret into exploration time steps (when the set $W_t$ is not empty), and exploitation time steps (when the set $W_t$ is empty). To bound the regret at exploration time periods we note that virtual time indexes could be expressed by

$$\tau_{k,t} = \sum_{s=1}^t \exp \left( \frac{\Delta^2}{c^2} \sum_{\tau=s}^t \eta_{k,\tau} \right),$$

and that the expected number of observations from arm $k$ due to exploration and information flows together is at most $\frac{\Delta^2}{c^2} \log \tau_{k,T} + 1$. Subtracting the number of auxiliary observations $\sum_{t=1}^T \eta_{k,t}$ one obtains
the first term in the upper bound. To analyze regret at exploitation time periods we use Chernoff-Hoeffding inequality to bound the probability that a sub-optimal arm would have the highest estimated reward, given the minimal number of observations that must be collected on each arm.

The upper bound in Theorem 2 holds for any arbitrary sample path of information arrivals that is captured by the matrix $H$, and matches the lower bound in Theorem 1 with respect to dependence on the sample path of information arrivals $\eta_{k,t}$'s, as well as the time horizon $T$, the number of arms $K$, and the minimum expected reward difference $\Delta$. This establishes a minimax regret rate of order

$$\sum_{k \in K} \log \left( \sum_{t=1}^T \exp \left( -c \cdot \sum_{s=1}^t \eta_{k,s} \right) \right)$$

for the MAB problem with information flows that is formulated here, where $c$ is a constant that may depend on problem parameters such as $K$, $\Delta$, and $\sigma$. Theorem 2 also implies that the adaptive exploration policy guarantees the best achievable regret (up to some multiplicative constant) under any arbitrary sample path of information arrivals. Notably, the optimality of the adaptive exploration policy applies to each of the settings that are described in §3.1, and matches the lower bounds that were established in §3.1.1 and §3.1.2 for any parametric values of $\lambda$ and $\kappa$.

**Corollary 1 (Near optimality under stationary information flows)** Let $\pi$ be an adaptive exploration policy with $c > 8$. If $\eta_{k,t}$'s are i.i.d. Bernoulli random variables with parameter $\lambda$ then, for every $T \geq 1$:

$$E_H[\mathcal{R}_S(H, T)] \leq \left( \sum_{k \in K} \Delta_k \right) \left( \frac{c \sigma^2}{\Delta^2} \log \left( \min \left\{ T + 1, \frac{c \sigma^2 + 10 \Delta^2}{\Delta^2 \lambda} \right\} \right) + C \right) \right),$$

for some absolute constant $C$.

**Corollary 2 (Near optimality under diminishing information flows)** Let $\pi$ be an adaptive exploration policy with $c > 8$. If $\eta_{k,t}$'s are random variables such that for some $\kappa \in \mathbb{R}^+$, $E[\eta_{k,s}] = \left\lceil \frac{\sigma \kappa^2}{2 \Delta^2} \log t \right\rceil$ for each arm $k \in K$ at each time step $t$, then, for every $T \geq 1$:

$$E_H[\mathcal{R}_S(H, T)] \leq \left( \sum_{k \in K} \Delta_k \right) \left( \frac{c \sigma^2}{\Delta^2} \log \left( 2 + \frac{T^{-1/2} - \frac{1}{\kappa^2}}{1 - \frac{\kappa}{4e}} + \frac{T^{1/2} - \frac{\kappa \sigma^2}{80 \Delta^2} - 1}{1 - \frac{\kappa \sigma^2}{20 \Delta^2}} \right) + C \right),$$

for some absolute constant $C$.

While the adaptive exploration policy can be used for achieving near optimal performance, it serves us mainly as a tool to communicate a broad approach for designing rate-optimal policies in the presence of unknown information flows: adjusting policies that are designed to achieve “good” performance in the absence of information flows, by endogenizing their exploration rates through virtual time indexes that are advanced whenever new information is revealed. Notably, the approach of regulating exploration rates based on realized information flows through advancing virtual time indexes (as specified in equation (1) and illustrated in Figure 1) can be applied quite broadly over various algorithmic approaches. In
the following section we demonstrate that rate-optimal performance may be achieved by applying this approach to known MAB policies that are rate optimal in the absence of auxiliary information flows.

5 Adjusting practical MAB policies

In §4 we introduced an approach to design efficient policies in the presence of auxiliary information by regulating the exploration rate of the policy using a virtual time index, and by advancing that virtual time through a properly selected multiplicative factor whenever auxiliary information is observed. To demonstrate the practicality of this approach, we next apply it to adjust the design of the $\epsilon$-greedy and UCB1 policies, that were shown to achieve rate-optimal performance in the classical MAB framework.

5.1 $\epsilon$-greedy with adaptive exploration

Consider the following adaptation of the $\epsilon$-greedy policy [Auer et al. 2002].

$\epsilon$-greedy with adaptive exploration. Input: a tuning parameter $c > 0$.

1. Initialization: set initial virtual times $\tau_{k,0} = 0$ for all $k \in K$

2. At each period $t = 1, 2, \ldots, T$:
   (a) Observe the vectors $\eta_t$, and $Z_t$
   (b) Update the virtual time steps for all $k \in K$:
   
   $\tau_{k,t} = \begin{cases} 
   \tau_{k,t-1} + 1 & \text{if } t < \left\lceil \frac{Kc\sigma^2}{\Delta^2} \right\rceil \\
   (\tau_{k,t-1} + 1) \cdot \exp \left( \frac{\eta_{k,t} \Delta^2}{c\sigma^2} \right) & \text{if } t = \left\lceil \frac{Kc\sigma^2}{\Delta^2} \right\rceil \\
   (\tau_{k,t-1} + 1) \cdot \exp \left( \frac{\eta_{k,t} \Delta^2}{c\sigma^2} \right) & \text{if } t > \left\lceil \frac{Kc\sigma^2}{\Delta^2} \right\rceil 
   \end{cases}$

   (c) With probability $\min \left\{ \frac{c\sigma^2}{\Delta^2} \sum_{k' = 1}^{K} \frac{1}{\tau_{k',t}}, 1 \right\}$ select an arm at random: (exploration)

   $\pi_t = k$ with probability $\frac{1}{\tau_{k,t}}$, for all $k \in K$

   Otherwise, select an arm with the highest estimated reward: (exploitation)

   $\pi_t = \arg \max_{k \in K} \hat{X}_{k,n_{k,t}}$

   (d) Receive and observe a reward $X_{\pi_t,t}$
The $\epsilon$-greedy with adaptive exploration policy chooses arms uniformly at random (explores) up to time step $t^* = \left\lceil \frac{Kc\sigma^2}{\Delta^2} \right\rceil$. Then, at period $t^*$ the policy advances the virtual time indexes associated with the different arms according to all the auxiliary information that has arrived up to time $t^*$. From period $t^* + 1$ and on, the policy advances the virtual time indexes based on auxiliary information that has arrived since the last period. Finally, at each step $t > t^*$, the policy explores with probability that is proportional to $\sum_{k'}^K \tau_{k',t}^{-1}$, and otherwise pulls the arm with the highest empirical mean reward. In exploration periods, arm $k$ is explored with a probability that is weighted by $\tau_{k,t}^{-1}$, based on the virtual time index that is associated with that arm. Notably, in the $\epsilon$-greedy with adaptive exploration policy virtual time indexes are advanced in the same manner that was detailed in §4 in the description of the adaptive exploration policy, except for an initial phase when for simplicity of analysis these indexes are set equal to the time index $t$. The following result characterizes the guaranteed performance and establishes the rate optimality of $\epsilon$-greedy with adaptive exploration in the presence of unknown information flows.

**Theorem 3 (Near optimality of $\epsilon$-greedy with adaptive exploration)** Let $\pi$ be an $\epsilon$-greedy with adaptive exploration policy tuned by $c > \max\left\{16, \frac{10\Delta^2}{\sigma^2}\right\}$. Then, there exists a time index $t^*$ such that for every $T > t^*$, and for any information arrival matrix $H$, one has:

$$R_{\pi}^S(H, T) \leq \sum_{k \in K} \Delta_k \left( \frac{C_6}{\Delta^2} \log \left( \sum_{t=t^*+1}^T \exp \left( \frac{-\Delta^2}{C_6} \sum_{s=1}^{t} \eta_{k,s} \right) \right) + C_7 \right),$$

where $C_6$, and $C_7$ are positive constants that depend only on $\sigma$.

**Key ideas in the proof.** The proof of Theorem 3 follows ideas that are similar to the ones described in the proof of Theorem 2, applying them to adjust the analysis of $\epsilon$-greedy policy in [Auer et al., 2002]. We decompose overall regret into exploration and exploitation time periods. To bound the regret at exploration time periods, we express the virtual times similarly to (2). Letting $t_m$ denote the time step at which the $m^{th}$ auxiliary observation for arm $k$ realizes. We establish an upper bound on the expected number of exploration time periods for arm $k$ in the time interval $[t_m, t_{m+1} - 1]$, which scales linearly with $\frac{c\sigma^2}{\Delta^2} \log \left( \frac{t_{m+1}}{t_m} \right) - 1$. Summing over all possible values of $m$, we obtain that the regret over exploration time periods is bounded from above by

$$\sum_{k \in K} \Delta_k \cdot \frac{c\sigma^2}{\Delta^2} \log \left( \frac{e}{t^* - 1/2} \sum_{t=t^*+1}^T \exp \left( \frac{-\Delta^2}{c\sigma^2} \sum_{s=1}^{t} \eta_{k,s} \right) + e \right).$$

To analyze regret over exploitation time periods we first lower bound the number of observations of each arm using Bernstein inequality, and then apply Chernoff-Hoeffding inequality to bound the probability that a sub-optimal arm would have the highest estimated reward, given the minimal number
of observations on each arm. This regret component is constant term whenever $c > \max\left\{16, \frac{10\Delta^2}{\sigma^2}\right\}$.

### 5.2 UCB1 with adaptive exploration

Consider the following adaptation of the UCB1 policy ([Auer et al., 2002]).

**UCB1 with adaptive exploration.** Inputs: a tuning constant $c$, and estimated reward differences $\left\{\hat{\Delta}_k\right\}_{k \in \mathcal{K}}$.

1. Initialization: set initial virtual times $\tau_{k,0} = 0$ for all $k \in \mathcal{K}$

2. At each period $t = 1, \ldots, T$:
   (a) Observe the vectors $\eta_t$, and $Z_t$
   (b) Update the virtual times: $\tau_{k,t} = (\tau_{k,t-1} + 1) \cdot \exp\left(\frac{\eta_k\hat{\Delta}_k^2}{4c\sigma^2}\right)$ for all $k \in \mathcal{K}$
   (c) Select the arm $\pi_t = \begin{cases} t & \text{if } t \leq K \\ \arg\max_{k \in \mathcal{K}} \left\{ \bar{X}_{k,n_k,t} + \sqrt{\frac{\sigma^2 \log \tau_{k,t}}{n_k,t}} \right\} & \text{if } t > K \end{cases}$
   (d) Receive and observe a reward $X_{\pi_t,t}$

The UCB1 with adaptive exploration policy takes actions using the upper-confidence bounds $\bar{X}_{k,n_k,t} + \sqrt{\frac{\sigma^2 \log \tau_{k,t}}{n_k,t}}$, where virtual time indexes $\tau_{k,t}$ are advanced multiplicatively (using the parameters $\left\{\hat{\Delta}_k\right\}$) whenever new information is realized. Notably, even though there are no explicit exploration periods in UCB1, in the absence of auxiliary information flows this policy selects suboptimal actions at rate $1/t$, and therefore essentially explores at the same rate. Therefore, virtual times indexes is advanced essentially as in previously discussed policies that explicitly explores at rate $1/t$. The following result establishes the rate optimality of UCB1 with adaptive exploration in the presence of unknown information flows.

**Theorem 4** Let $\pi$ be the adaptive UCB1 algorithm run with $\hat{\Delta}_k = \Delta$ for all $k \in \mathcal{K}$. Then, for any $T \geq 1$, $K \geq 2$, and auxiliary information arrival matrix $H$:

$$R^*_S(H,T) \leq \sum_{k \in \mathcal{K}} \Delta_k \left( \frac{C_8}{\Delta^2} \log \left( \sum_{t=1}^T \exp\left(\frac{-\Delta^2}{C_8} \sum_{s=1}^{t-1} \eta_{k,s}\right) \right) + C_9 \right),$$

where $C_8$, and $C_9$ are positive constants that depend only on $\sigma$.

**Key ideas in the proof.** The proof adjusts the analysis of original UCB1 policy. Pulling a suboptimal arm $k$ at time step $t$ implies at least one of the following three possibilities: (i) the empirical average of
the best arm deviates from its mean; (ii) the empirical mean of arm $k$ deviates from its mean; or (iii) arm $k$ has not been pulled sufficiently often in the sense that

$$\tilde{n}_{k,t-1} \leq l_{k,t} - \sum_{s=1}^{t} \eta_{k,s},$$

where $l_{k,t} = \frac{4 \sigma^2 \log(\tau_{k,t})}{\Delta_k^2}$ and $\tilde{n}_{k,t-1}$ is the number of times arm $k$ is pulled up to time $t$. The probability of the first two events can be bounded using Chernoff-Hoeffding inequality, and the probability of the third one can be bounded using:

$$\sum_{t=1}^{T} \mathbb{1} \left\{ \pi_t = k, \tilde{n}_{k,t-1} \leq l_{k,t} - \sum_{s=1}^{t} \eta_{k,s} \right\} \leq \max_{1 \leq t \leq T} \left\{ l_{k,t} - \sum_{s=1}^{t} \eta_{k,s} \right\}.$$

Therefore, we establish that for any $\hat{\Delta}_k \leq \Delta_k$, one has

$$\mathcal{R}_T^D(H,T) \leq \sum_{k \in \mathcal{K} \setminus \{k^*\}} \Delta_k \left( \frac{C_8}{\Delta_k^2} \cdot \max_{1 \leq t \leq T} \log \left( \sum_{m=1}^{t} \exp \left( \frac{\Delta_k^2}{C_8 \sum_{s=m}^{t} \eta_{k,s} - \Delta_k^2 / C_8 \sum_{s=1}^{t} \eta_{k,s}} \right) \right) + C_9 \right).$$

plugging in $\hat{\Delta}_k = \Delta$, the result follows. ■

### 5.3 Numerical analysis

To demonstrate the practical value of the adaptive exploration approach and the validity of the performance bounds that were established in the previous sections we analyze the empirical performance policies in the presence of unknown information flows.

**Setup.** For each of the three reward distribution profiles that are listed in Table 1 (capturing three levels of problem “hardness”), we considered three information arrival processes: stationary information flows with $\eta_{k,t}$’s being i.i.d. Bernoulli random variables with mean $\lambda = 500/T$; diminishing information flows where $\sum_{s=1}^{t} \eta_{k,t} = \lfloor \frac{t}{4} \rfloor$ for each time period $t$ and for each arm $k$; and no auxiliary information flows, a basic setting that is used as a benchmark for the former two cases. We experimented with the three policies that have been discussed in the previous sections: Adaptive exploration, $\epsilon$-greedy with adaptive exploration, and UCB1 with adaptive exploration, with a variety of tuning parameters. For each of the reward profiles in Table 1 we tracked the average empirical regret $\sum_{t=1}^{T} (\mu^* - \mu_{\pi_t})$, that is, the average performance difference between the best arm and the policy, over a decision horizon of $T = 10^6$ periods. Averaging over 100 repetitions the outcome approximates the expected regret.

**Results and discussion.** Plots comparing the regret accumulation of the various algorithms for the three different reward profiles appear in Figures 2, 3, and 4 and least squares estimation of the linear
Table 1: Three profiles of mean rewards, each includes mean rewards of 10 arms

<table>
<thead>
<tr>
<th>Profile</th>
<th>Mean rewards of different arms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8</td>
</tr>
<tr>
<td>2</td>
<td>0.9 0.8 0.8 0.8 0.7 0.7 0.7 0.6 0.6 0.6</td>
</tr>
<tr>
<td>3</td>
<td>0.9 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.6</td>
</tr>
</tbody>
</table>

Table 2: Slope and intercept estimates for profile 1

<table>
<thead>
<tr>
<th>Policy</th>
<th>No auxiliary information flows</th>
<th>Diminishing information flows</th>
<th>Stationary information flows</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>slope</td>
<td>intercept</td>
<td>slope</td>
</tr>
<tr>
<td>Adaptive exploration</td>
<td>17.98</td>
<td>0.59</td>
<td>14.65</td>
</tr>
<tr>
<td>Adaptive ϵ-greedy</td>
<td>17.92</td>
<td>−69.43</td>
<td>14.42</td>
</tr>
<tr>
<td>Adaptive UCB1</td>
<td>20.72</td>
<td>−63.67</td>
<td>18.84</td>
</tr>
</tbody>
</table>

Table 3: Slope and intercept estimates for profile 2

<table>
<thead>
<tr>
<th>Policy</th>
<th>No auxiliary information flows</th>
<th>Diminishing information flows</th>
<th>Stationary information flows</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>slope</td>
<td>intercept</td>
<td>slope</td>
</tr>
<tr>
<td>Adaptive exploration</td>
<td>36.00</td>
<td>0.61</td>
<td>29.43</td>
</tr>
<tr>
<td>Adaptive ϵ-greedy</td>
<td>35.79</td>
<td>−144.01</td>
<td>29.03</td>
</tr>
<tr>
<td>Adaptive UCB1</td>
<td>12.69</td>
<td>−30.51</td>
<td>7.37</td>
</tr>
</tbody>
</table>

Table 4: Slope and intercept estimates for profile 3

<table>
<thead>
<tr>
<th>Policy</th>
<th>No auxiliary information flows</th>
<th>Diminishing information flows</th>
<th>Stationary information flows</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>slope</td>
<td>intercept</td>
<td>slope</td>
</tr>
<tr>
<td>Adaptive exploration</td>
<td>5.98</td>
<td>2.32</td>
<td>2.60</td>
</tr>
<tr>
<td>Adaptive ϵ-greedy</td>
<td>5.88</td>
<td>−2.51</td>
<td>2.58</td>
</tr>
<tr>
<td>Adaptive UCB1</td>
<td>6.58</td>
<td>−8.30</td>
<td>3.75</td>
</tr>
</tbody>
</table>

Comparing the results under diminishing information flows with those without any auxiliary information flows verifies the analysis in §3.1.2. One may observe that diminishing information flows lead to

5A selection of $c = 0.2$ achieved the best performance in the case without auxiliary information flows, and in that sense this selection guarantees no loss relative to the best empirical tuning of the adjusted policies in the case without information flows. While achieving good performance, the value $c = 0.2$ does not belong to the range of parametric values for which the theoretical performance bounds were established. This is aligned with observations in the MAB literature where performance of tuning parameters belonging to a parametric region performance guarantees is dominated by the performance of some parametric values that do not belong to this region; see, e.g., related discussions in Auer et al. (2002).
lower accumulated regret, but the rate of regret is logarithmic rather than constant since information arrivals are not sufficiently frequent. Comparing the results with stationary information flows to those without any auxiliary information flow validates the analysis in §3.1.1, demonstrating that if the rate of information arrival, $\lambda$, is sufficiently high then the cumulative regret will be bounded from above by a constant that depends on the information arrival rate $\lambda$. More precisely, we observe that for the stationary information flows, the regret of the adaptive policies asymptotically converge to a constant that is independent of $T$.

In addition to validating performance bounds, the results demonstrate the practical value that may be captured by following the adjusted exploration approach in the presence of unknown information flows. Our performance bounds show that by appropriately utilizing available information flows that are sufficiently rich, one may reduce regret rates to a constant that is independent of $T$ (as is apparent from observing the left sides of Figures 2, 3, and 4 as well as observing the slope estimates in Tables 2, 3, and 4 for the stationary information flows). However, even when information flows are not rich enough to
allow reduced regret rates, utilizing this information may have big impact on the empirical performance in terms of reducing the accumulated regret and the multiplicative constant in the regret (as is apparent from comparing accumulated regret values and the reduced slopes in the middle parts relative to the left parts of Figures 2, 3 and 4, as well as comparing the values in Tables 2, 3 and 4, for the respective information flows). Such regret reduction is very valuable in addressing many practical problems for which MAB policies have been applied.

6 Reactive information flows

So far we have analyzed the performance of policies when information arrival processes are unknown, but fixed. In particular, information flows were assumed to be independent of the decision trajectory of the policy. In this section, we address some potential implications of endogenous information flows by considering a simple extension of our model in which information flows depend on the past decisions of the policy. For the sake of concreteness, we assume that the information flow on each arm is polynomially proportional (decreasing or increasing) to the number of times various arms were selected. For some global parameters $\gamma > 0$ and $\omega \geq 0$ that are fixed over arms and time periods, we assume that for each time step $t$ and arm $k$, the number of auxiliary observations received up to time $t$ on arm $k$ can be described as follows:

$$\sum_{s=1}^{t} \eta_{k,s} = \left( \rho_k \cdot \tilde{n}_{k,t-1}^\omega + \sum_{j \in K \setminus \{k\}} \alpha_{k,j} \tilde{n}_{j,t-1}^\gamma \right),$$

where $\tilde{n}_{k,t} = \sum_{s=1}^{t} 1\{\pi_s = k\}$ is the number of times arm $k$ is selected up to time $t$, $\rho_k \geq 0$ captures the dependence of auxiliary observations of arm $k$ on the past selections of arm $k$, and for each $j \neq k$ the parameter $\alpha_{k,j} \geq 0$ capture the dependence of auxiliary observations of arm $k$ on the past selections of arm $j$. We assume that there exist non-negative values $\bar{\rho}$, $\underline{\rho}$, $\bar{\alpha}$, and $\underline{\alpha}$ such that $\underline{\rho} \leq \rho_k \leq \bar{\rho}$ and
\[ \alpha \leq \alpha_{k,j} \leq \bar{\alpha} \text{ for all arms } k, j \text{ in } K. \]

While the structure of (3) introduces some limitation on the impact the decision path of the policy may have on the information flows, it still captures many types of dependencies. For example, when \( \gamma = 0 \), information arrivals are decoupled across arms, in the sense that selecting an action at a given period can impact only future information on that action. On the other hand, when \( \gamma > 0 \) a selection of a certain action may impact future information arrivals on all actions.

A key driver in the regret complexity of this MAB formulation with endogenous information flows is the order of the total number of times an arm is observed, \( n_{k,t} = \tilde{n}_{k,t} - 1 + \sum_{s=1}^{t} \eta_{k,t} \), relative to the order of the number of times that arm is pulled, \( \tilde{n}_{k,t} - 1 \). Therefore, a first observation is that as long as \( n_{k,t} \) and \( \tilde{n}_{k,t} - 1 \) are of the same order, one recovers the classical regret rates that appear in the stationary MAB literature; in particular, this is the case whenever \( \omega < 1 \). Therefore, in the rest of this section we focus on the case of \( \omega \geq 1 \).

A second observation is that when pulling an arm increases information arrival rates to other arms (that is, whenever \( \gamma > 0 \) and \( \bar{\alpha} > 0 \)), constant regret is achievable, and a myopic policy can guarantee rate-optimal performance. This observation is formalized by the following proposition.

**Proposition 1** Let \( \pi \) be a myopic policy that for the first \( K \) periods pulls each arm once, and at each later period pulls the arm with the highest estimated mean reward, while randomizing to break ties. Assume that \( \bar{\alpha} > 0 \) and \( \gamma > 0 \). Then, for any horizon length \( T \geq 1 \) and for any history-dependent information arrival matrix \( H \) such that (3) holds, one has

\[
R_\pi^\gamma(H, T) \leq C_{10} \cdot \Gamma(\tilde{\gamma}) \cdot \tilde{\gamma}^{\frac{\gamma}{2} - 1},
\]

where \( \Gamma(\cdot) \) is the gamma function, \( \tilde{\gamma} = \min\{\gamma, 1\} \), and \( C_{10} > 0 \) is a constant that depends on \( \bar{\alpha} \).

We next turn to characterize the case in which information flows are decoupled across arms, in the sense that selecting a certain arm does not impact information flows associated with other arms. To evaluate the performance of the adaptive exploration approach under this class of reactive information flows we use the adaptive exploration policy given in §4. We note that the adjusted policies presented in §5 could be shown to maintain similar performance guarantees.

**Theorem 5** (Near Optimality under decoupled endogenous information flows) Assume that \( \gamma = 0 \) and \( \omega \geq 1 \). Then:

(i) For any \( T \geq 1 \) and history-dependent information arrival matrix \( H \) that satisfies (3), the regret
incurred by any admissible policy \( \pi \in \mathcal{P} \), is bounded below as follows

\[
R_\pi^S(H, T) \geq \frac{C_{11} K}{\rho \Delta_k^{\frac{2}{\omega}} - 1} \left( \log \left( \frac{T}{K} \right) \right)^{1/\omega},
\]

for some positive constant \( C_{11} > 0 \) that depends on \( \sigma \).

(ii) Let \( \pi \) be the adaptive exploration policy (detailed in §4), tuned by \( c > 8 \). Then, for any \( T \geq \left\lceil \frac{4cK\sigma^2}{\Delta^2} \right\rceil \)

\[
R_\pi^S(H, T) \leq C_{12} \cdot \left( \log T \right)^{1/\omega} \cdot \sum_{k \in \mathcal{K} \setminus \{k^*\}} \frac{1}{\rho \Delta_k^{\frac{2}{\omega}} - 1} + \sum_{k \in \mathcal{K} \setminus \{k^*\}} C_{13} \Delta_k,
\]

where \( C_{12} \) and \( C_{13} \) are positive constants that depend on \( \sigma \) and \( c \).

Theorem 5 introduces matching lower and upper bounds that establish optimality under the class of decoupled endogenous information flows defined by (3) with \( \gamma = 0 \) and \( \omega \geq 1 \). For example, Theorem 5 implies that under our class of reactive information flows, whenever mean rewards are separated (that is, for each \( k \) one has that \( \Delta_k \) is independent of \( T \)), the best achievable performance is a regret is of order \( (\log T)^{1/\omega} \).

**Key ideas in the proof.** The first part of the result is derived by replacing the sum \( \sum_{s=1}^{t} \eta_{k,s} \) by its upper bound, \( \bar{\rho} n_{k,t-1}^\omega \), and following the proof of Theorem 1. To establish the second part of the result, we rely on the proof of Theorem 2 and perform a worst-case analysis for the part of the regret that is caused by exploration. We show that in this case, the \( j \)th exploration of arm \( k \) are separated by \( \frac{c\sigma^2}{\Delta^2} \tau_{k,t,k,j} \) time periods from other explorations of that arm, where \( t_{k,j} \) is the time step at which the \( j \)th exploration of arm \( k \) took place. The result is then obtained by plugging (3) into the expression of \( \tau_{k,t} \) and observing that \( t_{k,j} - t_{k,j-1} \leq T \) for all possible \( j \).

Notably, in proving the second part of Theorem 5 we assume no prior knowledge of either the class of information arrival processes given by (3), or the parametric values under this structure. Therefore, the adaptive exploration policy is agnostic to the structure of the information flows.

7 Concluding remarks

**Summary of results.** In this paper we considered a generalization of the stationary multi-armed bandits problem in the presence of unknown and arbitrary information flows on each arm. We studied the impact of such auxiliary information on the design of efficient learning policies and on the performance that can be achieved. In particular, we studied the design of learning policies that adapt in real time to the unknown information arrival process. We introduced a general approach of adjusting existing MAB policies that are designed to perform well in the absence of auxiliary information flows by controlling
the exploration rate that is governed by the policy through advancing a virtual time indexes that are
customized for each arm every time information on this arm arrives. We established that using this
approach one may guarantee the best performance (in terms of minimax regret) that is achievable under
prior knowledge on the information arrival process, and demonstrated through performance bounds as
well as empirical experiments the effectiveness of this approach in adapting “of the shelf” MAB polices.

**Inseparable mean rewards.** As mentioned in §2.1, all our results hold also when \( \Delta \), the lower bound
on the minimum difference of expected rewards relative to the best arm, is allowed to scale with \( T \), and
in particular, to decrease with \( T \). When \( \Delta \) is of order \( T^{-1/2} \), this leads to a spectrum of minimax regret
rates, from order \( T^{1/2} \) in the absence of information flows (coinciding with the worst case of inseparable
mean rewards in the classical stochastic MAB framework), and constant regret that is independent of \( T \).

**Avenues for future research.** As mentioned in §2, our results hold when there exists some invertible
mapping \( \phi \) such that the auxiliary information \( Y \) has the same distribution as \( \phi(X) \). One interesting
direction for future research is to study cases where the mapping \( \phi \) is a priori unknown; such case may
require more sophisticated estimation of mean rewards relative to the empirical mean that was used
in this paper. One possible approach for addressing this challenging problem might be to estimate
the mapping \( \phi \) throughout some initial period, form an estimated inverse mapping \( \hat{\phi}^{-1} \), and only then
leverage auxiliary information flows for improving the estimators of the mean rewards.

In §6 we considered information flows that are allowed to be reactive to the history of decisions
and observations, and established that optimality of our approach for a class of such information
flows. Nevertheless, the characterization of minimax regret rates and optimal policy design for general
endogenous information flows is a challenging open problem.
A Proofs of Main results

A.1 Preliminaries

In all the following proofs, we will have the following notations: Let $\tilde{n}_{k,t}$ be the number of times arm $k$ is pulled by the policy up to time $t$ (the time steps that the output of the arm is observed due to free information arrival are excluded):

$$\tilde{n}_{k,t} = \sum_{s=1}^{t} 1\{\pi_s = k\}.$$ 

Let $n_{k,t}$ be the number of samples observed from arm $k$ up to time $t$:

$$n_{k,t} = \tilde{n}_{k,t} - 1 + \sum_{s=1}^{t} \eta_{k,s}.$$ 

Denote by $\bar{X}_{k,n_{k,t}}$ the empirical average of the $k$th arm reward after $n_{k,t}$ observations:

$$\bar{X}_{k,n_{k,t}} = \frac{\eta_{k,t} Y_{k,t} + \sum_{s=1}^{t-1} (\eta_{k,s} Y_{k,s} + 1\{\pi_s = k\} X_{k,s})}{n_{k,t}}.$$ 

For any policy $\pi$, and profile $\nu$, let $P^\pi_{\nu}$, $E^\pi_{\nu}$, and $R^\pi_{\nu}$ denote the probability, expectation, and regret when the arm rewards are distributed according to $\nu$.

A.2 Proof of Theorem

Step 1 (Notations and definitions). We define the distribution profiles $\nu^{(m,q)}$ with $m, q \in \{1, \ldots, K\}$ as follows:

$$\nu^{(m,q)}_k = \begin{cases} 
N(0, \sigma^2) & \text{if } k = m \\
N(+\Delta, \sigma^2) & \text{if } k = q \neq m \\
N(-\Delta, \sigma^2) & \text{o.w.}
\end{cases}.$$ 

For example one has that,

$$\nu^{(1,1)} = \begin{pmatrix} 
N(0, \sigma^2) \\
N(-\Delta, \sigma^2) \\
\vdots \\
N(-\Delta, \sigma^2)
\end{pmatrix}, \quad \nu^{(1,2)} = \begin{pmatrix} 
N(0, \sigma^2) \\
N(+\Delta, \sigma^2) \\
\vdots \\
N(-\Delta, \sigma^2)
\end{pmatrix}, \quad \nu^{(1,3)} = \begin{pmatrix} 
N(0, \sigma^2) \\
N(-\Delta, \sigma^2)
\end{pmatrix}, \ldots, \nu^{(1,K)} = \begin{pmatrix} 
N(0, \sigma^2) \\
N(-\Delta, \sigma^2)
\end{pmatrix}.$$
Step 2 (Lower bound decomposition). We note that
\[ R_\pi^\pi(H, T) \geq \max_{m, q \in \{1, \ldots, K\}} \{ R_\pi^\nu(m, q)(H, T) \} \geq \frac{1}{K} \sum_{m=1}^{K} \max_{q \in \{1, \ldots, K\}} \{ R_\pi^\nu(m, q)(H, T) \}. \] (4)

Step 3 (A naive lower bound for \( \max_{q \in \{1, \ldots, K\}} \{ R_\pi^\nu(m, q)(H, T) \} \)). We note that
\[ \max_{q \in \{1, \ldots, K\}} \{ R_\pi^\nu(m, q)(H, T) \} \geq R_\pi^\nu(m, m)(H, T) = \Delta \cdot \sum_{k \in K \setminus \{m\}} \mathbb{E}_\nu(m, m)[\tilde{n}_k, T]. \] (5)

Step 4 (An information theoretic lower bound). For any profile \( \nu \), denote by \( \nu_t \) the distribution of the observed rewards up to time \( t \) under \( \nu \). By Lemma 2, for any \( q \neq m \), one has
\[ \text{KL}(\nu_t^{(m,m)}, \nu_t^{(m,q)}) = \frac{2\Delta^2}{\sigma^2} \cdot \mathbb{E}_\nu(m, m)[\tilde{n}_{q,t}] = \frac{2\Delta^2}{\sigma^2} \left( \mathbb{E}_\nu(m, m)[\tilde{n}_{q,t-1}] + \sum_{s=1}^{t} \eta_{q,s} \right). \] (6)

One obtains:
\[
\max_{q \in \{1, \ldots, K\}} \{ R_\pi^{\nu(m,q)}(H, T) \} \geq \frac{1}{K} R_\pi^{\nu(m,m)}(H, T) + \frac{1}{K} \sum_{q \in K \setminus \{m\}} R_\pi^{\nu(m,q)}(H, T)
\geq \frac{\Delta}{K} \sum_{t=1}^{T} \sum_{k \in K \setminus \{m\}} \mathbb{P}_\nu(m, m)\{\pi_t = k\} + \frac{\Delta}{K} \sum_{q \in K \setminus \{m\}} \sum_{t=1}^{T} \mathbb{P}_\nu(m, q)\{\pi_t = m\}
= \frac{\Delta}{K} \sum_{t=1}^{T} \sum_{q \in K \setminus \{m\}} \left( \mathbb{P}_\nu(m, m)\{\pi_t = q\} + \mathbb{P}_\nu(m, q)\{\pi_t = m\} \right)
\geq \frac{\Delta}{2K} \sum_{t=1}^{T} \sum_{q \in K \setminus \{m\}} \exp(-\text{KL}(\nu_t^{(m,m)}, \nu_t^{(m,q)}))
\geq \frac{\Delta}{2K} \sum_{t=1}^{T} \sum_{q \in K \setminus \{m\}} \exp \left[ -\frac{2\Delta^2}{\sigma^2} \left( \mathbb{E}_\nu(m, m)[\tilde{n}_{q,t-1}] + \sum_{s=1}^{t} \eta_{q,s} \right) \right]
\geq \frac{\Delta}{2K} \sum_{t=1}^{T} \sum_{q \in K \setminus \{m\}} \exp \left[ -\frac{2\Delta^2}{\sigma^2} \left( \mathbb{E}_\nu(m, m)[\tilde{n}_{q,T}] + \sum_{s=1}^{t} \eta_{q,s} \right) \right]
= \sum_{q \in K \setminus \{m\}} \frac{\Delta}{2K} \exp \left( -\frac{2\Delta^2}{\sigma^2} \cdot \mathbb{E}_\nu(m, m)[\tilde{n}_{q,T}] \right) \sum_{t=1}^{T} \exp \left( -\frac{2\Delta^2}{\sigma^2} \sum_{s=1}^{t} \eta_{q,s} \right), \] (7)

where (a) follows from Lemma 1, (b) holds by (6), and (c) follows from \( \tilde{n}_{q,T} \geq \tilde{n}_{q,t-1} \) for \( t \in T \).
Step 5 (Unifying the lower bounds in steps 3 and 4). Using (5), and (7), we establish:

\[
\max_{q \in \{1, \ldots, K\}} \left\{ R_{\pi_{\nu_{(m,q)}}}^\nu(H, T) \right\} \geq \frac{\Delta}{2} \sum_{k \in K \setminus \{m\}} \left( E_{\nu_{\pi_{\nu_{(m,m)}}}}[\tilde{n}_{k,T}] + \exp \left( -\frac{2\Delta^2}{\sigma^2} \cdot \frac{\sum_{t=1}^T \exp \left( -\frac{2\Delta^2}{\sigma^2} \cdot \eta_{k,s} \right)}{2K} \right) \sum_{t=1}^T \exp \left( -\frac{2\Delta^2}{\sigma^2} \cdot \eta_{k,s} \right) \right) \]
\[
\geq \frac{\Delta}{2} \sum_{k \in K \setminus \{m\}} \min_{x > 0} \left( x + \frac{\exp \left( -\frac{2\Delta^2}{\sigma^2} \cdot x \right)}{2K} \sum_{t=1}^T \exp \left( -\frac{2\Delta^2}{\sigma^2} \cdot \eta_{k,s} \right) \right) \]
\[
\geq \frac{\sigma^2}{4\Delta} \sum_{k \in K \setminus \{m\}} \log \left( \frac{\Delta^2}{\sigma^2 K} \sum_{t=1}^T \exp \left( -\frac{2\Delta^2}{\sigma^2} \cdot \eta_{k,s} \right) \right), \tag{8}
\]

where (a) follows from \( x + \gamma e^{-\kappa x} \geq \frac{\log \gamma \kappa}{\kappa} \) for \( \gamma, \kappa, x > 0 \). (Note that the function \( x + \gamma e^{-\kappa x} \geq \frac{\log \gamma \kappa}{\kappa} \) is a convex function and we can find its minimum by taking its derivative and putting it equal to zero.) The result is then established by putting together (4), and (8). ■

A.3 Proof of Theorem 2

Step 1 (Decomposing the regret). Fix \( T \geq 1, K \geq 2, \) and \( H \in \{0, 1\}^{K \times T} \), and profile \( \nu \). For simplicity of notation, define events \( E_{k,t} \) that at time \( t \), and for arm \( k \) we have

\[
n_{k,t} \geq c \frac{\sigma^2}{\Delta^2} \log (\tau_{k,t}).
\]

In words, this event means at time \( t \) the policy has seen arm \( k \) an enough number of times and will not perform exploration on this arm at this time step. Let \( \pi \) be the Adaptive-exploration policy. We decompose the regret of a suboptimal arm \( k \neq k^* \):

\[
\Delta_k \cdot E_{\pi}^\nu \left[ \sum_{t=1}^T \mathbb{1} \{ \pi_t = k \} \right] \leq \Delta_k \cdot E_{\pi}^\nu \left[ \sum_{t=1}^T \mathbb{1} \left\{ k = \arg \min_{j \in W_t} n_{j,t} \wedge E_{k,t} \right\} \right]
\leq \Delta_k \cdot E_{\pi}^\nu \left[ \sum_{t=1}^T \mathbb{1} \left\{ k = \arg \max_{j \in K} \tilde{X}_{j,n_{k,t}} \wedge \bigcap_{j \in K} E_{j,t} \right\} \right], \tag{9}
\]

where \( E_{k,t} \) is the complement of \( E_{k,t} \). The component \( J_{k,1} \) is the expected number of times arm \( k \) is pulled due to exploration, while the the component \( J_{k,2} \) is the one due to exploitation.
Step 2 (Analysis of $\tau_{k,t}$). We will later find upper bounds for the quantities $J_{k,1}$ and $J_{k,2}$ that are functions of the virtual time $\tau_{k,T}$. A simple induction results in the following expression:

$$\tau_{k,t} = \sum_{s=1}^{t} \exp \left( \frac{\Delta^2}{c\sigma^2} \sum_{\tau=s}^{t} \eta_{k,\tau} \right).$$

(10)

Step 3 (Analysis of $J_{k,1}$). Define

$$\mathcal{E}_{k,t} = \eta_{k,t} + \sum_{s=1}^{t-1} (\eta_{k,s} + 1 \left\{ k = \arg \min_{j \in W_s} n_{j,s} \land \bar{E}_{k,s} \right\}).$$

(11)

By induction, we next show that for $t = 1, \ldots, T + 1$:

$$\mathcal{E}_{k,t} \leq \frac{c\sigma^2}{\Delta^2} \log (\tau_{k,t}) + 1.$$

(12)

The base case is

$$\mathcal{E}_{k,1} = \eta_{k,1} = \frac{c\sigma^2}{\Delta^2} \log \left( \exp \left( \frac{\Delta^2}{c\sigma^2} \eta_{k,1} \right) \right) = \frac{c\sigma^2}{\Delta^2} \log (\tau_{k,1}) \leq \frac{c\sigma^2}{\Delta^2} \log (\tau_{k,1}) + 1.$$

Assume that (12) holds for all $s \leq t$. To establish that it also holds for $t + 1$, we will show that

$$\mathcal{E}_{k,t+1} \leq \frac{c\sigma^2}{\Delta^2} \log (\tau_{k,t}) + 1 + \eta_{k,t+1}.$$

(13)

We consider two cases. In the first one, $1 \left\{ k = \arg \min_{j \in W_s} n_{j,s} \land \bar{E}_{k,s} \right\} = 0$. As a result, we have

$$\mathcal{E}_{k,t+1} = \mathcal{E}_{k,t} + \eta_{k,t+1} \leq \frac{c\sigma^2}{\Delta^2} \log (\tau_{k,t}) + 1 + \eta_{k,t+1}.$$

In the second case, $1 \left\{ k = \arg \min_{j \in W_s} n_{j,t} \land \bar{E}_{k,t} \right\} = 1$. This implies $\mathcal{E}_{k,t} \leq n_{k,t} < \frac{c\sigma^2}{\Delta^2} \log (\tau_{k,t})$. Therefore,

$$\mathcal{E}_{k,t+1} = \mathcal{E}_{k,t} + \eta_{k,t+1} + 1 \leq \frac{c\sigma^2}{\Delta^2} \log (\tau_{k,t}) + 1 + \eta_{k,t+1},$$

and therefore (13) is established. The following inequality completes the induction:

$$\frac{c\sigma^2}{\Delta^2} \log (\tau_{k,t}) + 1 + \eta_{k,t+1} = \frac{c\sigma^2}{\Delta^2} \log \left( \tau_{k,t} \cdot \exp \left( \frac{\Delta^2}{c\sigma^2} \eta_{k,t+1} \right) \right) + 1 \leq \frac{c\sigma^2}{\Delta^2} \log (\tau_{k,t+1}) + 1.$$
Therefore, one obtains:

\[
J_{k,1} \overset{(a)}{=} \mathbb{E}_{\nu}^\tau [\mathcal{E}_{k,T+1}] - \sum_{t=1}^{T} \eta_{k,t}
\]

\[
\overset{(b)}{\leq} \frac{c\sigma^2}{\Delta^2} \log (\tau_{k,T+1}) + 1 - \sum_{t=1}^{T} \eta_{k,t}
\]

\[
= \frac{c\sigma^2}{\Delta^2} \log (\tau_{k,T+1}) + 1 + \frac{c\sigma^2}{\Delta^2} \log \left( \exp \left( -\frac{\Delta^2}{c\sigma^2} \sum_{t=1}^{T} \eta_{k,t} \right) \right)
\]

\[
\overset{(c)}{\leq} \frac{c\sigma^2}{\Delta^2} \log \left( 1 + \sum_{t=1}^{T} \exp \left( -\frac{\Delta^2}{c\sigma^2} \sum_{s=1}^{t} \eta_{k,s} \right) \right) + 1,
\]

(14)

where (a) and (b) follow from (11) and (12) respectively, and (c) holds by replacing \(\tau_{k,T+1}\) with the expression in (10).

**Step 4 (Analysis of \(J_{k,2}\)).** We note that:

\[
J_{k,2} \leq \sum_{t=1}^{T} \mathbb{P} \left\{ \bar{X}_{k,n_{k,t}} > \bar{X}_{k^*,n_{k^*,t}} \wedge E_{k,t} \right\}.
\]

(15)

We upper bound each summand as follows:

\[
\mathbb{P} \left\{ \bar{X}_{k,n_{k,t}} > \bar{X}_{k^*,n_{k,t}} \wedge E_{k,t} \right\} \leq \mathbb{P} \left\{ \bar{X}_{k,n_{k,t}} > \mu_k + \frac{\Delta_k}{2} \wedge E_{k,t} \right\}
\]

\[
+ \mathbb{P} \left\{ \bar{X}_{k^*,n_{k,t}} < \mu_{k^*} + \frac{\Delta_k}{2} \wedge E_{k,t} \right\}.
\]

(16)

We next upper bound the first term on the right-hand side of the above inequality; the second term can be treated similarly. Now, one obtains:

\[
\mathbb{P} \left\{ \bar{X}_{k,n_{k,t}} > \mu_k + \frac{\Delta_k}{2} \wedge E_{k,t} \right\} \overset{(a)}{\leq} \mathbb{P} \left\{ \bar{X}_{k,n_{k,t}} > \mu_k + \frac{\Delta_k}{2} \mid E_{k,t} \right\}
\]

\[
\overset{(b)}{\leq} \exp \left( -\frac{\Delta^2}{8\Delta^2} \log \tau_{k,t} \right)
\]

\[
\leq \frac{1}{(\tau_{k,t})^{c/8}} \leq \frac{1}{t^{c/8}},
\]

(17)

where (a) holds by the conditional probability definition, (b) follows from Lemma 3 and the fact that \(n_{k,t} \geq n_{k^*,t}\), and (c) is a result of (10). The results is then established by putting together (9), (14), (15), (16), and (10). This concludes the proof.
A.4 Proof of Corollary 1

\[ \mathbb{E}_H [R_S^\pi(H, T)] \leq \mathbb{E}_H \left[ \sum_{k \in K} \frac{c^2 \Delta_k}{\Delta^2} \log \left( \sum_{t=1}^{T} \exp \left( \frac{-\Delta^2}{c^2} \sum_{s=1}^{t} \eta_{k,s} \right) \right) + C \Delta_k \right] \]

\[ \leq \sum_{k \in K} \left[ \frac{c^2 \Delta_k}{\Delta^2} \log \left( \mathbb{E}_H \left[ \sum_{t=1}^{T} \exp \left( \frac{-\Delta^2}{c^2} \sum_{s=1}^{t} \eta_{k,s} \right) \right] \right) + C \Delta_k \right] \]

\[ \leq \left( \sum_{k \in K} \Delta_k \right) \cdot \left( \frac{c^2 \Delta_k}{\Delta^2} \log \left( \frac{1 - \exp \left( -T \lambda / 10 \right)}{\lambda / 10} + \frac{1 - \exp \left( -\Delta^2 T \lambda / c^2 \right)}{\Delta^2 \lambda / c^2} \right) + C \right) \]

\[ \leq \left( \sum_{k \in K} \Delta_k \right) \cdot \left( \frac{c^2 \Delta_k}{\Delta^2} \log \left( \frac{1}{\lambda / 10} + \frac{1}{\Delta^2 \lambda / c^2} \right) + C \right), \quad (18) \]

where: (a) holds by Theorem 2 \( \text{(b)} \) follows from Jensen’s inequality and the concavity of \( \log(\cdot) \), and (c) holds by Lemmas 5 and 6. By noting that \( \sum_{t=0}^{T} \exp \left( -\Delta^2 / c^2 \sum_{s=1}^{t} \eta_{k,s} \right) < T + 1 \), the result is established. \( \blacksquare \)

A.5 Proof of Corollary 2

\[ \mathbb{E}_H [R_S^\pi(H, T)] \leq \mathbb{E}_H \left[ \sum_{k \in K} \frac{c^2 \Delta_k}{\Delta^2} \log \left( \sum_{t=1}^{T} \exp \left( \frac{-\Delta^2}{c^2} \sum_{s=1}^{t} \eta_{k,s} \right) \right) + C \Delta_k \right] \]

\[ \leq \sum_{k \in K} \left[ \frac{c^2 \Delta_k}{\Delta^2} \log \left( \mathbb{E}_H \left[ \sum_{t=1}^{T} \exp \left( \frac{-\Delta^2}{c^2} \sum_{s=1}^{t} \eta_{k,s} \right) \right] \right) + C \Delta_k \right] \]

\[ \leq \left( \sum_{k \in K} \Delta_k \right) \cdot \left( \frac{c^2 \Delta_k}{\Delta^2} \log \left( \sum_{t=1}^{T} \exp \left( \frac{-\kappa}{4c} \log t \right) + \exp \left( \frac{-\kappa^2}{20 \Delta^2} \log t \right) \right) + C \right) \]

\[ = \left( \sum_{k \in K} \Delta_k \right) \cdot \left( \frac{c^2 \Delta_k}{\Delta^2} \log \left( \sum_{t=1}^{T} \frac{t^{\frac{-\kappa}{4c}} + t^{-\kappa^2 / 20 \Delta^2}}{t^\frac{-\kappa}{4c} + t^{-\kappa^2 / 20 \Delta^2}} \right) + C \right) \]

\[ \leq \left( \sum_{k \in K} \Delta_k \right) \cdot \left( \frac{c^2 \Delta_k}{\Delta^2} \log \left( 2 + \int_{1}^{T} \frac{t^{\frac{-\kappa}{4c}} + t^{-\kappa^2 / 20 \Delta^2}}{t^\frac{-\kappa}{4c} + t^{-\kappa^2 / 20 \Delta^2}} \, dt \right) + C \right) \]

\[ = \left( \sum_{k \in K} \Delta_k \right) \cdot \left( \frac{c^2 \Delta_k}{\Delta^2} \log \left( 2 + \frac{T^{1 - \frac{\kappa}{4c}} - 1}{1 - \frac{\kappa}{4c}} + \frac{T^{1 - \frac{\kappa^2}{20 \Delta^2}} - 1}{1 - \frac{\kappa^2}{20 \Delta^2}} \right) + C \right) \quad (19) \]

where: (a) holds by Theorem 2 \( \text{(b)} \) follows from Jensen’s inequality and the concavity of \( \log(\cdot) \), and (c) holds by Lemma 5. \( \blacksquare \)

A.6 Proof of Theorem 3

The structure of this proof is very similar to that of the proof of Theorem 2. Let \( t^* = \lfloor \frac{K c^2 \Delta^2}{\lambda} \rfloor \), and fix \( T \geq t^*, K \geq 2, H \in \{0,1\}^{K \times T} \), and profile \( \nu \in S \).
Step 1 (Decomposing the regret). We decompose the regret of a suboptimal arm $k \neq k^*$:

$$\Delta_k \cdot \mathbb{E}_\pi^n \left[ \sum_{t=1}^T 1 \{ \pi_t = k \} \right] = \Delta_k \cdot \mathbb{E}_\pi^n \left[ \sum_{t=1}^T 1 \{ \pi_t = k, \text{ policy does exploration} \} \right]$$

$$+ \Delta_k \cdot \mathbb{E}_\pi^n \left[ \sum_{t=1}^T 1 \{ \pi_t = k, \text{ policy does exploitation} \} \right].$$

The component $J_{k,1}$ is the expected number of times arm $k$ is pulled due to exploration, while the the component $J_{k,2}$ is the one due to exploitation.

Step 2 (Analysis of $\tau_{k,t}$). We will later find upper bounds for the quantities $J_{k,1}$ and $J_{k,2}$ that are functions of the virtual time $\tau_{k,T}$. Assume $t \geq t^*$. A simple induction results in the following expression:

$$\tau_{k,t} = t^* \exp \left( \frac{\Delta^2}{c\sigma^2} \sum_{\tau=1}^t \eta_{k,\tau} \right) + \sum_{s=t^*+1}^t \exp \left( \frac{\Delta^2}{c\sigma^2} \sum_{\tau=s}^t \eta_{k,\tau} \right).$$

Step 3 (Analysis of $J_{k,1}$). Let $\bar{M} = \sum_{t=1}^T \eta_{k,t}$, and $\bar{M} = \sum_{t=1}^{t^*} \eta_{k,t}$. Let $t_m$ be the time step at which the $m^{th}$ auxiliary observation for arm $k$ is received, that is,

$$t_m = \begin{cases} 
\max \left\{ t^*, \min \left\{ 0 \leq t \leq T \mid \sum_{s=1}^t \eta_{k,s} = m \right\} \right\} & \text{if } M \leq m \leq \bar{M} \\
T + 1 & \text{if } m = \bar{M} + 1
\end{cases}.$$
Note that we dropped the index \( k \) in the definitions above for simplicity of notation. Also define 
\[
\tau_{k,T+1} = (\tau_{k,T} + 1) \cdot \exp \left( \frac{\eta_{k,T} \Delta^2}{c \sigma^2} \right).
\]
One has:

\[
J_{k,1} = \sum_{t=1}^{T} \min \left\{ \frac{c \sigma^2}{\Delta^2} \sum_{k'=1}^{K} \frac{1}{\tau_{k',t}}, 1 \right\} \cdot \frac{1}{\tau_{k,t}}
\]

\[
= \frac{t^* - 1}{K} + \sum_{t=t^*}^{T} \frac{c \sigma^2}{\Delta^2 \tau_{k,t}}
\]

\[
= \frac{t^* - 1}{K} + \sum_{m=M}^{M} \sum_{s=0}^{t_{m+1} - t_m - 1} \frac{c \sigma^2}{\Delta^2 \tau_{k,t,m} + s}
\]

\[
\leq \frac{t^* - 1}{K} + \sum_{m=M}^{M} \frac{c \sigma^2}{\Delta^2} \log \left( \frac{\tau_{k,t,m} + t_{m+1} - t_m - 1 + 1/2}{\tau_{k,t,m} - 1/2} \right)
\]

\[
\leq \frac{t^* - 1}{K} + \sum_{m=M}^{M} \frac{c \sigma^2}{\Delta^2} \log \left( \frac{\exp \left( \frac{-\Delta^2}{c \sigma^2} \right) \cdot \left( \tau_{k,t,m+1} - 1/2 \right)}{\tau_{k,t,m} - 1/2} \right)
\]

\[
= \frac{t^* - 1}{K} + \frac{c \sigma^2}{\Delta^2} \log \left( \frac{\tau_{k,t^*+1} - 1/2}{\tau_{k,t^*} - 1/2} \right) - \sum_{t=t^*+1}^{T} \eta_{k,t}
\]

\[
= \frac{t^* - 1}{K} + \frac{c \sigma^2}{\Delta^2} \log \left( \frac{\tau_{k,T+1} - 1/2}{\tau_{k,t^*} - 1/2} \right) - \sum_{t=t^*+1}^{T} \eta_{k,t}
\]

\[
\leq \frac{t^* - 1}{K} + \frac{c \sigma^2}{\Delta^2} \log \left( \frac{t^* - 1/2}{t^* - 1/2} \right) - \sum_{t=1}^{T} \eta_{k,t}
\]

\[
\leq \frac{t^* - 1}{K} + \frac{c \sigma^2}{\Delta^2} \log \left( \frac{1}{t^* - 1/2} \sum_{t=t^*+1}^{T} \exp \left( \frac{-\Delta^2}{c \sigma^2} \sum_{s=1}^{t} \eta_{k,s} \right) \right) + 1
\]

\[
\leq \frac{c \sigma^2}{\Delta^2} \log \left( \frac{e}{t^* - 1/2} \sum_{t=t^*+1}^{T} \exp \left( \frac{-\Delta^2}{c \sigma^2} \sum_{s=1}^{t} \eta_{k,s} \right) \right) + e,
\]

where (a), (b), and (c) follow from Lemma 7, the fact that 
\[
\frac{\tau_{k,t,m} + t_{m+1} - t_m - 1 + 1/2}{\tau_{k,t,m} - 1/2}
\]

and \( 21 \), respectively.

**Step 4 (Analysis of \( n_{k,t} \)).** To analyze \( J_{k,2} \) we bound \( n_{k,t} \), the number of samples of arm \( k \) at each time \( t \) such that \( t \geq t^* \). Fix \( t \geq t^* \). Then, \( n_{k,t} \) is a summation of independent Bernoulli r.v.’s. One has

\[
E \left[ n_{k,t} \right] = \frac{t^* - 1}{K} + \sum_{s=1}^{t} \eta_{k,s} + \sum_{s=t^* + 1}^{t-1} \frac{c \sigma^2}{\Delta^2 \tau_{k,s}}.
\]

(23)
The term \( \sum_{s=t^*}^{t-1} \frac{c\sigma^2}{\Delta^2 \tau_{k,s}} \) can be bounded similar to Step 3. Let \( \hat{M}_t = \sum_{s=1}^{t-1} \eta_{k,s} \) and \( M = \sum_{s=1}^{t^*} \eta_{k,s} \). Let \( t_{m,t} \) be the time step at which the \( m^{th} \) auxiliary observation for arm \( k \) is received up to time \( t - 1 \), that is,

\[
t_{m,t} = \begin{cases} 
\max \left\{ t^*, \min \left\{ 0 \leq s \leq T \mid \sum_{s=1}^{q} \eta_{k,q} = m \right\} \right\} & \text{if } M \leq m \leq \hat{M}_t \\
 t - 1 & \text{if } m = \hat{M}_t + 1
\end{cases}
\]

One has that

\[
\sum_{s=t^*}^{t-1} \frac{c\sigma^2}{\Delta^2 \tau_{k,s}} = \sum_{m=M}^{\hat{M}_t} \sum_{s=0}^{t_{m+1,t} - t_{m,t} - 1} \frac{c\sigma^2}{\Delta^2 (\tau_{k,m,t} + s)} \geq \sum_{m=M}^{\hat{M}_t} \frac{c\sigma^2}{\Delta^2} \log \left( \frac{\tau_{k,t_{m+1,t}}}{\tau_{k,t_{m,t}}} \right) - \sum_{s=t^*+1}^{t-1} \eta_{k,s} = c \frac{\sigma^2}{\Delta^2} \log \left( \frac{\tau_{k,t-1}}{\tau_{k,t^*}} \right) - \sum_{s=1}^{t-1} \eta_{k,s} \tag{24}
\]

where (a) follows from Lemma 7. Putting together (23) and (24), one has

\[
\mathbb{E}[n_{k,t}] \geq \frac{t^* - 1}{K} + c \frac{\sigma^2}{\Delta^2} \log \left( \frac{\tau_{k,t-1}}{t^*} \right) - \frac{1}{K} + c \frac{\sigma^2}{\Delta^2} \log \left( \frac{e \tau_{k,t-1}}{t^*} \right) = x_t. \tag{25}
\]

Now by Lemma 4, we have

\[
\mathbb{P}\left\{ n_{k,t} \leq \frac{x_t}{2} \right\} \leq \exp \left( -\frac{x_t}{10} \right) = \exp \left( -\frac{1}{10K} \right) \cdot \left( \frac{t^*}{e \tau_{k,t-1}} \right)^{-\frac{c\sigma^2}{100\Delta^2}} \leq \exp \left( \frac{1}{10K} \right) \cdot \left( \frac{t^*}{e (t - 1)} \right)^{-\frac{c\sigma^2}{100\Delta^2}}, \tag{26}
\]

where the last inequality follows from (21).

**Step 5 (Analysis of \( J_{k,2} \)).** We note that:

\[
J_{k,2} \leq \sum_{t=t^*}^{T} \mathbb{P}\left\{ \hat{X}_{k,n_{k,t}} > X_{k^*,n_{k^*,t}} \right\}. \tag{27}
\]
We upper bound each summand as follows:

\[
\mathbb{P}\left\{ \bar{X}_{k,n_k,t} > \bar{X}_{k^*,n_k^*,t} \right\} \leq \mathbb{P}\left\{ \bar{X}_{k,n_k,t} > \mu_k + \frac{\Delta_k}{2} \right\} + \mathbb{P}\left\{ \bar{X}_{k^*,n_{k^*},t} < \mu_k + \frac{\Delta_k}{2} \right\}.
\]

We next bound the first term on the right-hand side of the above inequality; the second term can be treated similarly. One obtains:

\[
\mathbb{P}\left\{ \bar{X}_{k,n_k,t} > \mu_k + \frac{\Delta_k}{2} \right\} \leq \mathbb{P}\left\{ \bar{X}_{k,n_k,t} > \mu_k + \frac{\Delta_k}{2} \mid n_{k,t} > \frac{x_t}{2} \right\} + \mathbb{P}\left\{ n_{k,t} \leq \frac{x_t}{2} \right\}
\]

\[
\leq \exp\left( \frac{-\Delta^2_k x_t}{16\sigma^2} \right) + \exp\left( \frac{1}{10K} \cdot \left( \frac{t^*}{e(t-1)} \right)^{\frac{c\sigma^2}{10\Delta^2}} \right)
\]

\[
\leq \exp\left( \frac{\Delta^2_k}{16K\sigma^2} \right) \cdot \exp\left( \frac{-c}{16} \log \left( \frac{e\tau_{k,t}}{t^*} \right) \right) + \exp\left( \frac{1}{10K} \cdot \left( \frac{t^*}{e(t-1)} \right)^{\frac{c\sigma^2}{10\Delta^2}} \right)
\]

\[
\leq \exp\left( \frac{\Delta^2_k}{16K\sigma^2} \right) \cdot \left( \frac{t^*}{e(t-1)} \right)^{\frac{c\sigma^2}{10\Delta^2}} + \exp\left( \frac{1}{10K} \cdot \left( \frac{t^*}{e(t-1)} \right)^{\frac{c\sigma^2}{10\Delta^2}} \right),
\]

where: (a) holds by the conditional probability definition; (b) follows from Lemma 3 together with (26); and (c) holds by (21). Putting together (20), (22), (27), (28), and (29), the result is established.

\[\text{(29)}\]

A.7 Proof of Theorem 4

Fix a profile \(\nu \in \mathcal{S}\) and consider a suboptimal arm \(k \neq k^*\). If arm \(k\) is pulled at time \(t\), then

\[
\bar{X}_{k,n_k,t} + \sqrt{\frac{c\sigma^2 \log \tau_{k,t}}{n_{k,t}}} \geq \bar{X}_{k^*,n_{k^*},t} + \sqrt{\frac{c\sigma^2 \log \tau_{k^*,t}}{n_{k^*,t}}}.
\]

Therefore, at least one of the following three events must occur:

\[
E_{1,t} = \left\{ \bar{X}_{k,n_k,t} \geq \mu_k + \sqrt{\frac{c\sigma^2 \log \tau_{k,t}}{n_{k,t}}} \right\}, \quad E_{2,t} = \left\{ \bar{X}_{k^*,n_{k^*},t} \leq \mu_{k^*} - \sqrt{\frac{c\sigma^2 \log \tau_{k^*,t}}{n_{k^*,t}}} \right\},
\]

\[
E_{3,t} = \left\{ \Delta_k \leq 2\sqrt{\frac{c\sigma^2 \log \tau_{k,t}}{n_{k,t}}} \right\}.
\]

To see why this is true, assume that all the above events fail then, we have

\[
\bar{X}_{k,n_k,t} + \sqrt{\frac{c\sigma^2 \log \tau_{k,t}}{n_{k,t}}} < \mu_k + 2\sqrt{\frac{c\sigma^2 \log \tau_{k,t}}{n_{k,t}}} < \mu_k + \Delta_k = \mu_{k^*} < \bar{X}_{k^*,n_{k^*},t} + \sqrt{\frac{c\sigma^2 \log \tau_{k^*,t}}{n_{k^*,t}}}.
\]
For any sequence \( \{l_{k,t}\}_{t \in \mathcal{T}} \), and \( \{\hat{l}_{k,t}\}_{t \in \mathcal{T}} \), such that \( \hat{l}_{k,t} \geq l_{k,t} \) for all \( t \in \mathcal{T} \), we can see that

\[
\mathbb{E}_\nu[\hat{n}_{k,T}] = \mathbb{E}_\nu \left[ \sum_{t=1}^{T} \mathbb{1} \{ \pi_t = k \} \right] = \mathbb{E}_\nu \left[ \sum_{t=1}^{T} \mathbb{1} \{ \pi_t = k, n_{k,t} \leq l_{k,t} \} + \mathbb{1} \{ \pi_t = k, n_{k,t} > l_{k,t} \} \right]
\]

\[
\leq \mathbb{E}_\nu \left[ \sum_{t=1}^{T} \mathbb{1} \{ \pi_t = k, \tilde{n}_{k,t-1} \leq \hat{l}_{k,t} - \sum_{s=1}^{t} \eta_{k,s} \} \right] + \sum_{t=1}^{T} \mathbb{P} \{ \pi_t = k, n_{k,t} > l_{k,t} \}
\]

\[
\leq \max_{1 \leq t \leq T} \left\{ \hat{l}_{k,t} - \sum_{s=1}^{t} \eta_{k,s} \right\} + \sum_{t=1}^{T} \mathbb{P} \{ \pi_t = k, n_{k,t} > l_{k,t} \}
\]

If we set

\[
l_{k,t} = \frac{4c\sigma^2 \log \left( \frac{l_{k,t}}{\Delta_k^2} \right)}{\Delta_k^2}, \quad \text{and} \quad \hat{l}_{k,t} = \frac{4c\sigma^2 \log \left( \frac{l_{k,t}}{\Delta_k^2} \right)}{\Delta_k^2}
\]

then, to have \( n_{k,t} > l_{k,t} \), it must be the case that \( E_{3,t} \) does not occur. Therefore,

\[
\mathbb{E}_\nu[\hat{n}_{k,T}] \leq \max_{1 \leq t \leq T} \left\{ \hat{l}_{k,t} - \sum_{s=1}^{t} \eta_{k,s} \right\} + \sum_{t=1}^{T} \mathbb{P} \{ \pi_t = k, E_{3,t} \text{ fails} \}
\]

\[
\leq \max_{1 \leq t \leq T} \left\{ \hat{l}_{k,t} - \sum_{s=1}^{t} \eta_{k,s} \right\} + \sum_{t=1}^{T} \mathbb{P} \{ E_{1,t} \text{ or } E_{2,t} \}
\]

\[
\leq \max_{1 \leq t \leq T} \left\{ \hat{l}_{k,t} - \sum_{s=1}^{t} \eta_{k,s} \right\} + \sum_{t=1}^{T} \frac{2}{c/2},
\]

where the last inequality follows from Lemma 3 and union bound. Using an induction, one may obtain that

\[
\tau_{k,t} = \sum_{s=1}^{t} \exp \left( \frac{\hat{\Delta}_k^2}{4c\sigma^2} \sum_{\tau=s}^{t} \eta_{k,\tau} \right) \geq t.
\]

Plugging in this expression into (31), one obtains

\[
\mathbb{E}_\nu[\hat{n}_{k,T}] \leq \max_{1 \leq t \leq T} \left\{ \frac{4c\sigma^2}{\hat{\Delta}_k^2} \log \left( \sum_{s=1}^{t} \exp \left( \frac{-\hat{\Delta}_k^2}{4c\sigma^2} \sum_{\tau=s}^{t} \eta_{k,\tau} \right) \right) \right\} + \sum_{t=1}^{T} \frac{2}{\tau_{k,t}}
\]

\[
= \frac{4c\sigma^2}{\Delta_k^2} \log \left( \sum_{t=1}^{T} \exp \left( \frac{-\Delta_k^2}{4c\sigma^2} \sum_{s=1}^{t-1} \eta_{k,s} \right) \right) + \sum_{t=1}^{T} \frac{2}{\tau_{k,t}}
\]

\[
\leq \frac{4c\sigma^2}{\Delta_k^2} \log \left( \sum_{t=1}^{T} \exp \left( \frac{-\Delta_k^2}{4c\sigma^2} \sum_{s=1}^{t-1} \eta_{k,s} \right) \right) + \sum_{t=1}^{T} \frac{2}{t^{c/2}}
\]

Putting \( \hat{\Delta}_k = \Delta_k \) concludes the proof.
A.8 Proof of Theorem 5

First, we prove the lower bound. The first three steps of the proof of Theorem 1 holds for the new setting as well. In step 4, equation (6) can be modified as follows:

\[
\text{KL}(\nu_t^{(m,m)}, \nu_t^{(m,q)}) = \frac{2\Delta^2}{\sigma^2} \mathbb{E}_{\nu_t^{(m,m)}} [n_{q,t}] = \frac{2\Delta^2}{\sigma^2} \mathbb{E}_{\nu_t^{(m,m)}} \left[ \tilde{n}_{q,t} + \sum_{s=1}^{t} \eta_{q,s} \right] \leq \frac{2\Delta^2}{\sigma^2} \mathbb{E}_{\nu_t^{(m,m)}} [(1 + \tilde{\rho})\tilde{n}_{q,t-1}] ,
\]

where the last inequality follows from \( \tilde{n}_{q,t-1} \geq \tilde{n}_{q,t-1} \) for \( \omega \geq 1 \). Accordingly, (7) can be modified as well:

\[
\max_{q \in \{1, \ldots, K\}} \{ R_{\nu_t^{(m,q)}}(H, T) \} \geq \frac{\Delta^2}{2} \sum_{k \in K \backslash \{m\}} \mathbb{E}_{\nu_t^{(m,m)}} [\tilde{n}_{k,T}] + \frac{T}{2K} \mathbb{E}_{\nu_t^{(m,m)}} [(1 + \tilde{\rho})\tilde{n}_{q,T}] ,
\]

where \( \tilde{\rho} \) is obtained by setting \( \gamma, \kappa > 0 \), and \( \omega \geq 1 \).

Now, we analyze the regret of the adaptive exploration policy. Similar to the proof of the regret lower bound, we will modify the proof of Theorem 2 to make it work for the setting of §6. Steps 1, 2, and 4 of the mentioned proof hold for this setting as well. We will focus on adjusting step 3 in the proof, that is, analyzing \( J_{k,1} \). Define

\[
\tilde{J}_{k,1} := \sum_{t=1}^{T} 1 \left\{ k = \arg \min_{j \in W_t} n_{j,t} \wedge \tilde{E}_{k,t} \right\} ,
\]

By the structure of the adaptive exploration policy, if \( t \geq t^* := \frac{4cK\sigma^2}{\Delta^2} \) then

\[
\tilde{n}_{k,t} \geq \frac{c\sigma^2}{\Delta^2} \log (\tau_{k,t}) - \sum_{s=1}^{t} \eta_{k,s} .
\]
Note that by the definition of \( \tau_{k,t} \)'s, we have

\[
\frac{c\sigma^2}{\Delta^2} \log(\tau_{k,t}) - \frac{c\sigma^2}{\Delta^2} \log(\tau_{k,t-1}) = \frac{c\sigma^2}{\Delta^2} \log(1 + \frac{1}{\tau_{k,t-1}}) + \eta_{k,t},
\]

which implies

\[
\hat{n}_{k,t} \geq \frac{c\sigma^2}{\Delta^2} \log (\tau_{k,t}) - \sum_{s=1}^{t} \eta_{k,s} = \sum_{s=2}^{t} \frac{c\sigma^2}{\Delta^2} \log(1 + \frac{1}{\tau_{k,s-1}}).
\]  \( (36) \)

Let \( t_{k,j} \) be the time step at which the \( j \)th exploration is performed on arm \( k \). By the structure of the policy, and (36), for \( t_{k,j} \geq t^* \), we have

\[
\sum_{s=t_{k,j}}^{t_{k,j}+1-1} \frac{c\sigma^2}{\Delta^2} \log(1 + \frac{1}{\tau_{k,s-1}}) \geq 1.
\]

Note that \( \tau_{k,t} \) is a decreasing sequence, which implies

\[
t_{k,j+1} - t_{k,j} \geq \frac{c\sigma^2}{\Delta^2} \log(1 + \frac{1}{\tau_{k,j+1}}) \geq \frac{c\sigma^2}{\Delta^2} \tau_{k,t_{k,j}} \geq \frac{c\sigma^2}{\Delta^2} \exp \left( \frac{\Delta^2 (\rho(j-1)^\omega - 1)}{c\sigma^2} \right),
\]  \( (37) \)

where (a) follows from \( \log(1 + x) \leq x \) for every \( x \geq 1 \), and (b) follows from (10). Let \( j = \hat{J}_{k,1} \) in (37):

\[
\frac{c\sigma^2}{\Delta^2} \exp \left( \frac{\Delta^2 (\rho(\hat{J}_{k,1}-1)^\omega - 1)}{c\sigma^2} \right) \leq t_{k,\hat{J}_{k,1}} - t_{k,\hat{J}_{k,1}-1} \leq T - t^* \leq T,
\]

which implies

\[
\hat{J}_{k,1} \leq 1 + \left( \frac{1}{\rho} + \frac{c\sigma^2}{\rho \Delta^2} (2 + \log T) \right)^{\frac{1}{\omega}}.
\]

This concludes the proof. \( \blacksquare \)

### A.9 Proof of Proposition 1

Define \( \bar{\gamma} = \min\{1, \gamma\} \), and \( \bar{\alpha} = \min\{1, \alpha\} \). First, we will show that for any arm \( k \), and at any time step \( t \), we have

\[
n_{k,t} \geq \bar{\alpha}(t - 1)^{\bar{\gamma}}.
\]  \( (38) \)
We will consider two cases. In the first one, assume $\gamma \geq 1$ then, we will have
\[
\begin{align*}
n_{k,t} &= \tilde{n}_{k,t-1} + \sum_{s=1}^{t} \eta_{k,s} \geq \tilde{n}_{k,t-1} + \sum_{j \in \mathcal{K} \setminus \{k\}} \alpha \tilde{n}_{j,t-1} \\
&\geq \tilde{n}_{k,t-1} + \sum_{j \in \mathcal{K} \setminus \{k\}} \alpha \tilde{n}_{j,t-1} \\
&\geq \min\{1, \alpha\} \sum_{j=1}^{K} \tilde{n}_{j,t-1} = \min\{1, \alpha\}(t-1) = \tilde{\alpha}(t-1)^\gamma.
\end{align*}
\]

In the second case, assume $\gamma < 1$. One obtains
\[
\begin{align*}
n_{k,t} &\geq \tilde{n}_{k,t-1} + \sum_{j \in \mathcal{K} \setminus \{k\}} \alpha \tilde{n}_{j,t-1} \\
&\geq \tilde{n}_{k,t-1} + \sum_{j \in \mathcal{K} \setminus \{k\}} \alpha \tilde{n}_{j,t-1} \\
&\geq \min\{1, \alpha\} \sum_{j=1}^{K} \tilde{n}_{j,t-1} = \tilde{\alpha}(t-1)^\gamma,
\end{align*}
\]
where the last step follows from the fact that $x^\gamma + y^\gamma \geq (x+y)^\gamma$ for $x, y \geq 0$, and $\gamma < 1$.

In order to upper bound the regret, we will use the usual decomposition
\[
R^\text{greedy}_S(H, T) = \sum_{t=1}^{T} \sum_{k \in \mathcal{K} \setminus \{k^*\}} \Delta_k \mathbb{P}\{\pi_t = k\}.
\]

We upper bound each summand as follows:
\[
\mathbb{P}\{\pi_t = k\} \leq \mathbb{P}\{\tilde{X}_{k,n,k,t} > \tilde{X}_{k^*,n,k,t}\} \leq \mathbb{P}\{\tilde{X}_{k,n,k,t} > \mu_k + \frac{\Delta_k}{2}\} + \mathbb{P}\{\tilde{X}_{k^*,n,k^*,t} < \mu_{k^*} - \frac{\Delta_k}{2}\}. 
\]

We next upper bound the first term on the right-hand side of the above inequality; the second term can be treated similarly. Now, one obtains:
\[
\mathbb{P}\{\tilde{X}_{k,n,k,t} > \mu_k + \frac{\Delta_k}{2}\} \leq \mathbb{E}\left[\exp\left(-\frac{\Delta_k^2}{16\sigma_k^2 n_{k,t}}\right)\right] \leq \exp\left(-\frac{\Delta_k^2}{8\sigma_k^2 \tilde{\alpha}(t-1)^\gamma}\right) \quad (40)
\]

where (a), and (b) hold by Lemma 3 and (38). Finally, note that
\[
\sum_{t=1}^{T} \mathbb{P}\{\pi_t = k\} \leq \sum_{t=1}^{T} 2 \exp\left(-\frac{\Delta_k^2}{8\sigma_k^2 \tilde{\alpha}(t-1)^\gamma}\right) \leq 2 + \frac{\Gamma\left(\frac{1}{\gamma}\right)}{\tilde{\alpha}} \left(\frac{8\sigma_k^2}{\Delta_k^2}\right)^{1/\gamma},
\]
where the last inequality follows form $\sum_{t=1}^{\infty} \exp(-t^\gamma) \leq \frac{\Gamma\left(\frac{1}{\gamma}\right)}{\gamma}$. \qed
B Auxiliary Lemmas

Lemma 1 (Tsybakov 2008, Lemma 2.6) Let $\rho_0, \rho_1$ be two probability distributions supported on some set $X$, with $\rho_0$ absolutely continuous with respect to $\rho_1$. Then for any measurable function $\Psi : X \to \{0, 1\}$, one has:

$$\mathbb{P}_{\rho_0}\{\Psi(X) = 1\} + \mathbb{P}_{\rho_1}\{\Psi(X) = 0\} \geq \frac{1}{2} \exp(-\text{KL}(\rho_0, \rho_1)).$$

Lemma 2 (Gerchinovitz and Lattimore 2016, Lemma 1) Consider two profiles $\nu$, and $\nu'$. Denote by $\nu_t (\nu'_t)$ the distribution of the observed rewards up to time $t$ under $\nu (\nu'$ respectively). Let $n_{k,t}$ be the number of times a sample from arm $k$ has been observed up to time $t$, that is $n_{k,t} = \eta_{k,t} + \sum_{s=1}^{t-1} (\eta_{k,s} + 1\{\pi_s = k\})$. For any policy $\pi$, we have

$$\text{KL}(\nu_t, \nu'_t) = \sum_{k=1}^{K} \mathbb{E}_\nu [n_{k,t}] \cdot \text{KL}(\nu_k, \nu'_k).$$

Lemma 3 (Chernoff-Hoeffding bound) Let $X_1, \ldots, X_n$ be random variables such that $X_t$ is a $\sigma^2$-sub-Gaussian random variable conditioned on $X_1, \ldots, X_{t-1}$ and $\mathbb{E}[X_t | X_1, \ldots, X_{t-1}] = \mu$. Let $S_n = X_1 + \cdots + X_n$. Then for all $a \geq 0$

$$\mathbb{P}\{S_n \geq \mu + a\} \leq e^{-\frac{a^2}{2n\sigma^2}}, \quad \text{and} \quad \mathbb{P}\{S_n \leq \mu - a\} \leq e^{-\frac{a^2}{2n\sigma^2}}.$$

Lemma 4 (Bernstein inequality) Let $X_1, \ldots, X_n$ be random variables with range $[0, 1]$ and

$$\sum_{t=1}^{n} \text{Var}[X_t | X_{t-1}, \ldots, X_1] = \sigma^2.$$

Let $S_n = X_1 + \cdots + X_n$. Then for all $a \geq 0$

$$\Pr\{S_n \geq \mathbb{E}[S_n] + a\} \leq \exp\left(-\frac{a^2/2}{\sigma^2 + a/2}\right).$$

Lemma 5 Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli random variable with parameters $p_1, \ldots, p_n$, respectively, then, for any $\kappa > 0$:

$$\mathbb{E}\left[ e^{-\kappa \sum_{j=1}^{n} X_j} \right] \leq e^{-\sum_{j=1}^{n} p_j/10} + e^{-\kappa \sum_{j=1}^{n} p_j/2}.$$
Proof. Define event $E$ to be $\sum_{j=1}^{n} X_j < \sum_{j=1}^{n} p_j/2$. By Lemma 4, we have:

$$P\{E\} \leq e^{-\sum_{j=1}^{n} p_j/10}. \quad (42)$$

By the law of total expectation:

$$\mathbb{E} \left[ e^{-\kappa \sum_{j=1}^{n} X_j} \right] \leq \mathbb{E} \left[ e^{-\kappa \sum_{j=1}^{n} X_j} \right] \cdot P\{E\} + \mathbb{E} \left[ e^{-\kappa \sum_{j=1}^{n} X_j} \mid \bar{E} \right] \cdot P\{\bar{E}\} \leq \mathbb{E} \left[ e^{-\kappa \sum_{j=1}^{n} X_j} \mid \bar{E} \right] \leq e^{-\sum_{j=1}^{n} p_j/10} + e^{-\kappa \sum_{j=1}^{n} p_j/2},$$

where (a) holds by (42).

Lemma 6 For any $\kappa > 0$, we have $\sum_{j=n_1}^{n_2} e^{-\kappa j} \leq \frac{1}{\kappa} (e^{-\kappa(n_1-1)} - e^{\kappa n_2}).$

Lemma 7 For any $t > 1/2$, and $n \in \{0, 1, 2, \ldots\}$, we have $\log \frac{t+n+1}{t} \leq \sum_{s=0}^{n} \frac{1}{t+s} \leq \log \frac{t+n+1/2}{t-1/2}.$

Lemma 8 (Pinsker’s inequality) Let $\rho_0, \rho_1$ be two probability distributions supported on some set $\mathbb{X}$, with $\rho_0$ absolutely continuous with respect to $\rho_1$ then,

$$\|\rho_0 - \rho_1\|_1 \leq \sqrt{\frac{1}{2} KL(\rho_0, \rho_1)},$$

where $\|\rho_0 - \rho_1\|_1 = \sup_{A \subset \mathbb{X}} |\rho_0(A) - \rho_1(A)|$ is the variational distance between $\rho_0$ and $\rho_1$. 

45
C Auxiliary proofs

C.1 Proof of Subsection 3.1.1

We note that:

\[
\mathbb{E}_H \left[ R_S^\pi (H, T) \right] \geq \mathbb{E}_H \left[ \frac{\sigma^2 (K - 1)}{4K \Delta} \sum_{k=1}^{K} \log \left( \frac{\Delta^2}{\sigma^2 K} \sum_{t=1}^{T} \frac{1}{\sum_{s=1}^{T} \eta_{s,k}} \right) \right] \tag{a}
\]

\[
\geq \frac{\sigma^2 (K - 1)}{4 \Delta} \log \left( \frac{\Delta^2}{\sigma^2 K} \sum_{t=0}^{T-1} e^{-\frac{2 \Delta^2}{\sigma^2 T \lambda}} \right) \tag{b}
\]

\[
= \frac{\sigma^2 (K - 1)}{4 \Delta} \log \left( \frac{\Delta^2}{\sigma^2 K} \cdot \frac{1 - e^{-\frac{2 \Delta^2}{\sigma^2 T \lambda}}}{1 - e^{-\frac{2 \Delta^2}{\sigma^2 \lambda}}} \right) \tag{c}
\]

\[
\geq \frac{\sigma^2 (K - 1)}{4 \Delta} \log \left( \frac{1 - e^{-\frac{2 \Delta^2}{\sigma^2 T \lambda}}}{2 \lambda} \right), \quad (43)
\]

where (a) holds by Theorem 1, (b) follows from the fact that log-sum-exp is a convex function (see (Boyd and Vandenberghe 2004, Example 3.1.5)), and (c) follows from \(1 - e^{-\frac{2 \Delta^2}{\sigma^2 \lambda}} \leq 2 \Delta^2 \lambda / \sigma^2\). Now we consider the following cases:

1. If \(2 \Delta^2 T \lambda \leq \sigma^2 / 2\) then, by inequality \(1 - e^{-x} \geq 2(1 - e^{-1/2}) x\) for \(0 \leq x \leq 1/2\), and (45), we have

\[
\mathbb{E}_H \left[ R_S^\pi (H, T) \right] \geq \frac{\sigma^2 (K - 1)}{4 \Delta} \log \left( \frac{(1 - e^{-1/2}) \Delta^2 T}{\sigma^2 K} \right).\]

2. If \(2 \Delta^2 T \lambda \geq \sigma^2 / 2\) then, by inequality \(1 - e^{-\frac{2 \Delta^2}{\sigma^2 T \lambda}} \geq 1 - e^{-1/2}\), and (45), we have

\[
\mathbb{E}_H \left[ R_S^\pi (H, T) \right] \geq \frac{\sigma^2 (K - 1)}{4 \Delta} \log \left( \frac{1 - e^{-1/2}}{2 \lambda} \right). \tag{44}
\]

C.2 Analysis of the myopic policy under the setting of Example 3.1.1

Assume that \(\pi\) is a myopic policy. Consider a suboptimal arm \(k\). One has

\[
P \{ \pi_t = k \} \leq P \left\{ \bar{X}_{k,n,t} > \bar{X}_{k^*,n_t} \right\} \leq P \left\{ \bar{X}_{k,n,t} > \mu_k + \frac{\Delta_k}{2} \right\} + P \left\{ \bar{X}_{k^*,n_t} < \mu_{k^*} + \frac{\Delta_{k^*}}{2} \right\}.
\]

We will upper bound \(P \left\{ \bar{X}_{k,n,t} > \mu_k + \frac{\Delta_k}{2} \right\}. \quad P \left\{ \bar{X}_{k^*,n_t} < \mu_{k^*} + \frac{\Delta_{k^*}}{2} \right\}\) can be upper bounded similarly.
\[
\mathbb{P}\left\{ X_{k,n,t} > \mu_k + \frac{\Delta_k}{2} \right\} \leq \mathbb{P}\left\{ X_{k,n,t} > \mu_k + \frac{\Delta_k}{2} \left| \sum_{s=1}^{t} \eta_{k,s} > \frac{\lambda t}{2} \right\} + \mathbb{P}\left\{ \sum_{s=1}^{t} \eta_{k,s} \leq \frac{\lambda t}{2} \right\} \overset{(a)}{=} e^{-\frac{\Delta_k^2 M}{16\sigma^2}} + e^{-\frac{M}{16}},
\]

where (a) follows from Lemma 3 and Lemma 4. As a result the cumulative regret is upper bounded by

\[
\sum_{k \in \mathcal{K} \setminus \{k^*\}} \left( 32\sigma^2 \frac{\Delta_k}{\lambda} + \frac{20\Delta_k}{\lambda} \right).
\]

Therefore, if \( \Delta \) is a constant independent of \( T \) then the regret is upper bounded by a constant.

C.3 Proof of Subsection 3.1.2

We note that:

\[
E_{H} \left[ R_{S}^{\pi}(H,T) \right] \overset{(a)}{\geq} E_{H} \left[ \frac{\sigma^2(K-1)}{4K\Delta} \sum_{k=1}^{K} \log \left( \frac{\Delta_k^2}{\sigma^2 K} \sum_{t=1}^{T} e^{-\frac{\Delta_k^2}{\sigma^2} \sum_{s=1}^{t} \eta_{s,k}} \right) \right]
\]

\[
\overset{(b)}{\geq} \frac{\sigma^2(K-1)}{4\Delta} \log \left( \frac{\Delta_k^2}{\sigma^2 K} \sum_{t=1}^{T} e^{-\kappa \log t} \right)
\]

\[
= \frac{\sigma^2(K-1)}{4\Delta} \log \left( \frac{\Delta_k^2}{\sigma^2 K} \sum_{t=1}^{T} t^{-\kappa} \right)
\]

\[
\overset{(c)}{\geq} \frac{\sigma^2(K-1)}{4\Delta} \log \left( \frac{\Delta_k^2}{\sigma^2 K} \int_{t=1}^{T+1} t^{-\kappa} dt \right)
\]

\[
= \frac{\sigma^2(K-1)}{4\Delta} \log \left( \frac{\Delta_k^2}{\sigma^2 K} \cdot \frac{(T+1)^{1-\kappa} - 1}{1-\kappa} \right), \quad (45)
\]

where (a) holds by Theorem 1 (b) follows from the fact that log-sum-exp is a convex function (see (Boyd and Vandenberghe 2004, Example 3.1.5)). □

C.4 Analysis of the myopic policy under the setting of Example 3.1.2

In this section we will shortly prove that if \( \sum_{s=1}^{t} \eta_{s,k} = \left[ \frac{\sigma^2_k}{2\Delta^2} \log t \right] \) for each arm \( k \) and time step \( t \) then a myopic policy achieves an asymptotic constant regret if \( \Delta_k^2 \kappa > 16\sigma^2 \). For any profile \( \nu \), we have:

\[
R_{\nu}^{\pi}(H,T) \leq \sum_{k \in \mathcal{K} \setminus \{k^*\}} \Delta_k \cdot \sum_{t=1}^{T} e^{-\frac{-\Delta_k^2 n_{k,t}}{8\sigma^2}} + \mathbb{E} e^{-\frac{-\Delta_k^2 n_{k^*t}}{8\sigma^2}}
\]

\[
\leq \sum_{k \in \mathcal{K} \setminus \{k^*\}} \Delta_k \cdot \sum_{t=1}^{T} \frac{-\Delta_k^2}{16\Delta^2} + \frac{-\Delta_k^2}{16\sigma^2} \leq \sum_{k \in \mathcal{K} \setminus \{k^*\}} \frac{32\Delta_k^2}{\kappa\Delta_k}.
\]

47
References


Caron, S. and S. Bhagat (2013). Mixing bandits: A recipe for improved cold-start recommendations in a social network. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, pp. 11. ACM.


Lee, K.-C., A. Jalali, and A. Dasdan (2013). Real time bid optimization with smooth budget delivery in online advertising. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, pp. 1. ACM.


