Adaptive Sequential Experiments with Unknown Information Flows

Yonatan Gur
Stanford University

Ahmadreza Momeni*
Stanford University

September 25, 2019

Abstract

When information is limited, online recommendation services aim to strike a balance between maximizing immediate payoffs based on available information, and acquiring new information that is essential for maximizing future payoffs. This trade-off is captured by the multi-armed bandit (MAB) framework that has been studied and applied for designing sequential experiments when at each time epoch a single observation is collected on the action that was selected at that epoch. However, in many cases, additional information may become available between decision epochs. We introduce a generalized MAB formulation in which auxiliary information may appear arbitrarily over time. By obtaining matching lower and upper bounds, we characterize the minimax complexity of this family of problems as a function of the information arrival process, and study how salient characteristics of this process impact policy design and achievable performance. We establish that while Thompson sampling and UCB policies leverage additional information naturally, other policies, such as εt-greedy, do not exhibit such robustness. We introduce a virtual time indexes method for dynamically controlling the exploration rate of policies, and apply it for designing εt-greedy-type policies that, without any prior knowledge on the information arrival process, attain the best performance (in terms of regret rate) that is achievable when the information arrival process is a priori known. We use data from a large media site to empirically analyze the value that may be captured in practice by leveraging auxiliary information flows for content recommendations. We further apply our method to a contextual bandit framework that allows for personalized recommendations using idiosyncratic consumer characteristics.

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

1 Introduction

1.1 Background and motivation

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 techniques that leverage information such as explicit and implicit user preferences, product consumption and popularity, as well as consumer ratings (see, e.g., Hill et al. 1995, Konstan et al. 1997, Breese et al. 1998). While effective when ample relevant information is available, these techniques

*The authors are grateful to Omar Besbes, Ramesh Johary, and Lawrence M. Wein for their valuable comments. An initial version of this work, including some preliminary results, appeared in [Gur and Momeni 2018]. Correspondence: ygur@stanford.edu, amomenis@stanford.edu.
Figure 1: Additional information in product recommendations. In the depicted scenario, consumers of type (A) sequentially arrive to an online retailer’s organic recommendation system to search for products that match the terms “headphones” and “bluetooth.” Based on these terms, the retailer can recommend one of two products: an old brand that was already recommended many times to consumers, and a new brand that was never recommended before. In parallel, consumers of type (B) arrive to the new brand’s product page directly from a search engine (e.g., Google) by searching for “headphones,” “bluetooth,” and “Denisy” (the name of the new brand).

Additional Information

Denisy
Old item New item

(A)
(B)

Recommendation System

old item new item

Figure 1: Additional information in product recommendations. In the depicted scenario, consumers of type (A) sequentially arrive to an online retailer’s organic recommendation system to search for products that match the terms “headphones” and “bluetooth.” Based on these terms, the retailer can recommend one of two products: an old brand that was already recommended many times to consumers, and a new brand that was never recommended before. In parallel, consumers of type (B) arrive to the new brand’s product page directly from a search engine (e.g., Google) by searching for “headphones,” “bluetooth,” and “Denisy” (the name of the new brand).

tend to under-perform when encountering 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.

In the presence of new products, a recommendation system needs to strike a balance between maximizing instantaneous performance indicators (such as revenue) and collecting valuable information that is essential for optimizing future recommendations. To illustrate this tradeoff (see Figure 1), consider a consumer (A) that is searching for a product using the organic recommendation system of an online retailer (e.g., Amazon). The consumer provides key product characteristics (e.g., “headphones” and “bluetooth”), and based on this description (and, perhaps, additional factors such as the browsing history of the consumer) the retailer recommends products to the consumer. While some of the candidate products that fit the desired description may have already been recommended many times to consumers, and the mean returns from recommending them are known, other candidate products might be new brands that were not recommended to consumers before. Evaluating the mean returns from recommending new products requires experimentation that is essential for improving future recommendations, but might be costly if consumers are less likely to purchase these new products.

A well-studied framework that captures this trade-off between new information acquisition (exploitation), 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 repeatedly chooses between different arms and receives a reward associated with the selected arm. In the stochastic formulation of this problem, rewards from each arm are assumed to be identically distributed and independent across trials and arms, reward distributions are a priori unknown to the agent, and the objective of the
agent is to maximize the cumulative return over a certain time horizon. Several MAB settings 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, the aforementioned studies consider settings where in each time period information is obtained only on items that are recommended at that period, and design recommendation policies that are predicated on this information structure. However, in many practical instances 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 its organic recommender system; see browsing data analysis in Sharma et al. (2015) that estimate that product recommendation systems are responsible for at most 30% of site traffic and revenue and that substantial fraction of remaining traffic arrives to product pages directly from search engines, as well as Besbes et al. (2016) that report similar findings for content recommendations in media sites. To illustrate an additional source of traffic and revenue information, let us revert to Figure 1 and consider a different consumer (B) that does not use the organic recommender system of the retailer, but rather arrives to a certain product page directly from a search engine (e.g., Google) by searching for that particular brand. While consumers (A) and (B) may have inherently different preferences, the actions taken by consumer (B) after arriving to the product page (e.g., whether or not she bought the product) can be informative about the returns from recommending that product to other consumers. This additional information could potentially be used to improve the performance of recommendation algorithms, especially when the browsing history that is associated with some of the products is short.

The above example illustrates that additional information often becomes available between decision epochs of recommender systems, and that this information might be essential for designing effective product recommendations. In comparison with the restrictive information collection process that is assumed in more classical MAB formulations, the availability of such auxiliary information (and even the prospect of obtaining it in the future) should impact the design of sequential experiments and learning policies. When additional information is available one may potentially obtain better estimators for the mean returns, and therefore may need to “sacrifice” less decisions for exploration. While this intuition suggests that exploration can be reduced in the presence of additional information, it is a priori not clear exactly how the design of sequential experiments should depend on the information flows.

Moreover, monitoring exploration rates 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, 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 auxiliary information flows, an important challenge is to adapt in real time to an a priori unknown information arrival process and adjust exploration rates 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 that 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 of auxiliary information flows? (ii) How should the design of sequential experiments and efficient decision-making policies change in the presence of auxiliary information flows? (iii) How are achievable performance and policy design affected by the characteristics of the information arrival process, such as the frequency of auxiliary 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?

While we mainly focus here on a product recommendation setting, our formulation and approach are relevant for a broad array of practical settings where bandit models have been applied for designing sequential experiments and where various types of additional information might be realized and leveraged over time. Examples of such settings include dynamic pricing (Bastani et al. 2019), retail management (Zhang et al. 2010), clinical trials (Bertsimas et al. 2016, Anderer et al. 2019), as well as many machine learning domains (see a list of relevant settings in the survey by Pan and Yang 2009).

1.2 Main contributions

The main contribution of this paper lies in (1) introducing a new, generalized MAB framework with unknown and arbitrary information flows; (2) characterizing the minimax regret complexity of this broad class of problems and, by doing so, providing a sharp criterion for identifying policies that are rate-optimal in the presence of auxiliary information flows; and (3) proposing rate-optimal adaptive policies that demonstrate how sequential experiments should be designed in settings characterized by unknown information arrival processes. More specifically, our contribution is along the following lines.

(1) Modeling. To capture the salient features of settings with auxiliary information arrivals we formulate a new class of MAB problems that generalizes the classical stochastic MAB framework, by relaxing strong assumptions that are typically imposed on the information collection process. Our formulation considers a broad class of distributions that are informative about mean rewards, and allows observations from these distributions to arrive at arbitrary and a priori unknown rates and times, and therefore captures a large variety of real-world phenomena, yet maintains mathematical tractability.
(2) **Analysis of achievable performance.** We establish lower bounds on the performance that is achievable by *any* non-anticipating policy, 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 (together with other problem characteristics such as the length of the problem horizon and the number of arms). This provides a yardstick for evaluating the performance of policies, and a sharp criterion for identifying rate-optimal ones. Our results identify 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 first establish that policies based on posterior sampling and upper confidence bounds may leverage additional information naturally, and in that sense, exhibit remarkable robustness with respect to the information arrival process. This significantly generalizes the class of information collection processes to which such policies have been applied thus far. Nevertheless, we observe that other policies such as $\varepsilon_t$-greedy, that are of practical importance, do not exhibit such robustness.

We introduce a novel *virtual time indexes method* for endogenizing the exploration rate of policies, and apply it for designing $\varepsilon_t$-greedy-type policies that, without any prior knowledge on the information arrival process, attain the best performance (in terms of regret rate) that is achievable when the information arrival process is a priori known. This “best of all worlds” type of guarantee implies that rate optimality is achieved *uniformly* over the general class of information arrival processes at hand (including the case with no additional information). The virtual time indexes method adjusts the exploration rate of the policy in real time based on the amount of information that becomes available over time, through replacing the time index used by the policy with virtual time indexes that are dynamically updated. Whenever auxiliary information on a certain arm arrives, the virtual time index associated with that arm is carefully advanced to effectively reduce the rate at which the policy is experimenting with that arm.

Our method is simple yet quite general, and can be applied to a broad array of problems, including ones that are well-grounded in the practice of product recommendations. First, we demonstrate this using data from a large media site by empirically analyzing the value that may be captured by leveraging auxiliary information flows for recommending content with unknown impact on the future browsing path of readers. In addition, we extend our framework to a more general contextual bandit setting. The use of contexts allows for personalized recommendations that leverage idiosyncratic characteristics of (i) consumers that use the recommendations; and (ii) auxiliary information, for example, in the form of actions taken by consumers that arrive to product pages directly from global search engines.
1.3 Related work

**Recommender systems.** An active stream of literature has been studying recommender systems, focusing on modelling and maintaining connections between users and products; see, e.g., (Ansari et al. 2000), the survey by (Adomavicius and Tuzhilin 2005), and a book by (Ricci et al. 2011). One key element that impacts the performance of recommender systems is the often limited data that is available. Focusing on the prominent information acquisition aspect of the problem, several studies (to which we referred earlier) have addressed sequential recommendation problems using a MAB framework where in each time period information is obtained only on items that are recommended at that period. Another approach is to identify and leverage additional sources of relevant information. Following that avenue, Farias and Li (2019) consider the problem of estimating user-item propensities, and propose a method to incorporate auxiliary data such as browsing and search histories to enhance the predictive power of recommender systems. While their work concerns with the impact of auxiliary information in an offline prediction context, our paper focuses on the impact of auxiliary information streams on the design, information acquisition, and appropriate exploration rate in a sequential experimentation framework.

**Multi-armed bandits.** Since its inception, the MAB framework has been adopted for studying a variety of applications including clinical trials (Zelen 1969), strategic pricing (Bergemann and Välimäki 1996), assortment selection (Caro and Gallien 2007), online auctions (Kleinberg and Leighton 2003), online advertising (Pandey et al. 2007), and product recommendations (Madani and DeCoste 2005, Li et al. 2010). For a comprehensive overview of MAB formulations we refer the readers to 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 and Cesa-Bianchi (2012) that cover the machine learning literature and the so-called adversarial setting. A sharp regret characterization for the more traditional formulation (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 policies such as $\epsilon_t$-greedy, UCB1, and Thompson sampling; see, e.g., Auer et al. (2002) and Agrawal and Goyal (2013a).

The MAB framework focuses on balancing exploration and exploitation, typically under very little assumptions on the distribution of rewards, but with restrictive assumptions on the future information collection process. Correspondingly, 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 that may become available between pulls, and that might be essential for achieving good performance in many practical settings. 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 achievable performance and policy design; and (ii) adapting to a priori unknown sample path of information arrivals in real time.

As alluded to above, few MAB settings allow more information to 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. These settings consider specific information structures that are typically known a priori to the agent, as opposed to our formulation where the information arrival process is arbitrary and a priori unknown. (In §6 we demonstrate that the policy design approach we advance here could be applied to contextual MAB settings as well.) Another example is the full-information adversarial MAB setting, where rewards are arbitrary and can even be selected by an adversary (Auer et al. 1995, Freund and Schapire 1997). In this setting, at each time period the agent observes the rewards generated by all the arms (not only the one that was selected). The adversarial nature of this setting makes it fundamentally different in terms of achievable performance, analysis, and policy design, from the stochastic formulation that is adopted in this paper.

Balancing and regulating exploration. Several papers have considered settings of dynamic optimization with partial information and distinguished between cases where myopic policies guarantee optimal performance, and cases where exploration is essential, in the sense that myopic policies may lead to incomplete learning and large losses; see, e.g., Araman and Caldentey (2009), Farias and Van Roy (2010), Harrison et al. (2012), and den Boer and Zwart (2013) for dynamic pricing without knowing the demand function, Huh and Rusmevichientong (2009) and Besbes and Muharremoglu (2013) for inventory management without knowing the demand distribution, and Lee et al. (2003) in the context of technology development. Bastani et al. (2017) consider the contextual MAB framework and show that if the distribution of contexts guarantees sufficient diversity, then exploration becomes unnecessary and greedy policies can leverage the natural exploration that is embedded in the information diversity to achieve asymptotic optimality. Relatedly, Woodroofe (1979) and Sarkar (1991) consider a Bayesian one-armed contextual MAB problem and show that a myopic policy is asymptotically optimal when the discount factor converges to one. Considering sequential recommendations to customers that make decisions based on the relevance of recommendations, Bastani et al. (2018) show that classical MAB policies may over-explore and propose proper modifications for those policies.

On the other hand, few papers have studied cases where exploration is not only essential but should be enhanced in order to maintain optimality. For example, Cesa-Bianchi et al. (2006) introduce a partial monitoring setting where after playing an arm the agent does not observe the incurred loss but only a limited signal about it, and show that such feedback structure requires higher exploration rates. Besbes et al. (2018) consider a general 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. Shah et al. (2018) consider a platform in which the preferences of arriving users may depend on the experience of previous users, show that in this setting classical MAB policies may under-explore, and propose a balanced-exploration approach that leads to 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 dynamically endogenizing the rate at which a decision-making policy explores 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 rate-optimality without prior knowledge on the information arrival process.

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 [2.2]

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 \( \mu_k \). We denote the profile of rewards at time \( t \) by \( \mathbf{X}_t = (X_{1,t}, \ldots, X_{K,t})^\top \) and the profile of mean rewards by \( \mathbf{\mu} = (\mu_1, \ldots, \mu_K)^\top \). We further denote by \( \nu = (\nu_1, \ldots, \nu_K)^\top \) the distribution of the rewards profile \( \mathbf{X}_t \).

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\(^1\)

\[
\mu^* = \max_{k \in K} \{\mu_k\}, \quad k^* = \arg \max_{k \in K} \mu_k.
\]

We denote by \( \Delta_k = \mu^* - \mu_k \) the difference between the expected reward of the best arm and the expected reward of arm \( k \), and by \( \Delta \) a lower bound such that \( 0 < \Delta \leq \min_{k \in K \setminus \{k^*\}} \Delta_k \).

**Auxiliary information flows.** Before each round \( t \), the agent may or may not observe auxiliary

---

1. 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 \( \mathbb{E} e^{\lambda (X - \mathbb{E} X)} \leq e^{\sigma^2 \lambda^2 / 2} \). This broad class of distributions includes, for instance, Gaussian random variables, as well as any random variable with a bounded support (if \( X \in [a, b] \) then \( X \) is \( (b-a)^2 / 4 \)-sub-Gaussian) such as Bernoulli random variables. In particular, if a random variable is \( \sigma^2 \)-sub-Gaussian, it is also \( \tilde{\sigma}^2 \)-sub-Gaussian for all \( \tilde{\sigma} > \sigma \).

2. For simplicity, when using the \( \arg \min \) and \( \arg \max \) operators we assume ties to be broken in favor of the smaller index.
We define the worst-case regret. Let \( \pi \) be a policy, performance, and regret. Let \( \Pi = \{ \pi \in P \} \) be the class of admissible policies denoted by \( \Pi \). Policies in \( \Pi \) 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 denote by \( S = S(\Delta, \sigma^2, \hat{\sigma}^2) \) the class that includes pairs of allowed reward distribution profiles \((\nu, \nu')\), and assume that the parameters of this class are known.

We evaluate the performance of a policy \( \pi \in \Pi \) 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:

\[
\mathcal{R}_{\mathcal{S}}(H, T) = \sup_{(\nu, \nu') \in \mathcal{S}} \mathbb{E}_{(\nu, \nu')}^{\mathcal{S}} \left[ \sum_{t=1}^{T} (\mu^* - \mu_{\pi_t}) \right],
\]

where the expectation \( \mathbb{E}_{(\nu, \nu')}^{\mathcal{S}}[\cdot] \) is taken with respect to the noisy rewards and noisy auxiliary observations, as well as to the policy’s actions (throughout the paper we will denote by \( \mathbb{P}_{(\nu, \nu')}, \mathbb{E}_{(\nu, \nu')}^{\mathcal{S}}, \mathcal{R}_{(\nu, \nu')} \) the probability, expectation, and regret when the arms are selected according to policy \( \pi \), rewards are distributed according to \( \nu \), and auxiliary observations are distributed according to \( \nu' \)). We note that regret can only be incurred for decision made in epochs \( t = 1, \ldots, T \); the main distinction relative to classical regret formulations is that in (1) the mappings \( \{\pi_t; t = 1, \ldots, T\} \) can be measurable with
respect to sigma fields that also include information that arrives between decision epochs, as captured by the matrix $H$. We denote by $R^*_S(H, T) = \inf_{\pi \in \mathcal{P}} R_S^{\pi}(H, T)$ the best achievable guaranteed performance: the minimal regret that can be guaranteed by an admissible policy $\pi \in \mathcal{P}$. In the following sections we study the magnitude of $R^*_S(H, T)$ as a function of the information arrival process $H$.

### 2.1 A simple example with linear mappings

A special case of information flows consists of observations of independent random variables that are linear mappings of rewards. That is, there exist vectors $(\alpha_1, \ldots, \alpha_K) \in \mathbb{R}^K$ and $(\beta_1, \ldots, \beta_K) \in \mathbb{R}^K$ such that for all $k$ and $t$,

$$E[\beta_k + \alpha_k Y_{k,t}] = \mu_k.$$

While simple, this class of information flows captures the essence of utilizing data that is available for maximizing conversions of product recommendations (for example, in terms of purchases of the recommended products). Consumers that search for product features through the retailer’s organic recommender system can be different from consumers that arrive to a specific product page from an external search engine. Nevertheless, the conversion rates that are associated with these different types of consumers are both affected by features such as the attractiveness of the product page, clarity of description, and product specifications. This implies that one could expect the two conversion rates to be correlated. Denoting by $X_{k,t}$ the random reward of the recommender system from displaying item $k$ at time $t$, one could correspondingly model the conversion of a consumer arriving to the product page directly from external search just before time $t$ as a random variables $Y_{k,t}$ such that $\beta_k + \alpha_k E[Y_{k,t}] = \mu_k$.

This approximation suggests a linear relationship between the conversion rate of consumers to which the product is recommended and the conversion rate of consumers arriving to the product page from search.

While this example is given mainly for the purpose of illustrating our framework, the practicality of using linear mappings for improving product recommendations will be demonstrated empirically in §4.3 using data from a large media site. Nevertheless, we note that this example abstracts away from problem features that might be prominent in some settings, such as the idiosyncratic consumer characteristics that may be available to the recommender system in real time, and the relation between different pools of consumers that may be more complicated than the linear relation used here. In §6 we apply our approach to a more general contextual bandits formulation that captures these problem features.

---

3 The challenge of adapting to unknown mappings $\{\phi_k\}$ is discussed in §7. Nevertheless, we note that in many practical settings these mappings can be evaluated based on available information on consumer characteristics, similar products, and consumer behavior. For example, several methods have been developed to estimate click-through rates using MAB-based methods (see, e.g., Madani and DeCoste 2005) and feature-based or semantic-based decomposition methods (see, e.g., Regelson and Fain 2006, Richardson et al. 2007, Dave and Varma 2010). Since click-through rates to product pages are typically low (and since clicks are prerequisites for conversions), and due to a common delay between clicks and conversions, click-through rates can often be estimated before conversion rates; see, e.g., Rosales et al. 2012, Lee et al. 2012, Lee et al. 2013, Chapelle 2014, and the empirical comparison of several recommendations methods in Zhang et al. 2014.
2.2 Generalizations and extensions

While our model adopts the stochastic MAB framework of Lai and Robbins (1985) as a baseline and extends it to allow for additional information flows, it can be applied to more general frameworks, including ones where connection between auxiliary observations and mean rewards is established through the parametric structure of the problem rather than through mappings \( \phi_k \). In §6 we demonstrate this by extending the model of Goldenshluger and Zeevi (2013) where mean rewards linearly depend on stochastic context vectors, to derive performance bounds in the presence of additional information flows.

For simplicity we assume that only one information arrival can occur between consecutive time periods for each arm (that is, \( h_{k,t} \in \{0,1\} \) for each time \( t \) and arm \( k \)). In fact, all our results hold with more than one information arrival per time step per arm (allowing any integer values for the entries \( \{h_{k,t}\} \)).

We focus on a setting where the information arrival process (the matrix \( H \)) is unknown, but exogenous and does not depend on the sequence of decisions. While fully characterizing the regret complexity when information flows may depend on the history is a challenging open problem, in §7.3 we study policy design under a broad class of information flows that are reactive to the past decisions of the policy.

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 stochastic MAB literature. Nevertheless, we do not make any explicit assumption on the separability of mean rewards and our analysis and results hold for the more general case where \( \Delta \) is a decreasing function of the horizon length \( T \).

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

\[
R^\pi_{S}(H, T) \geq \frac{C_1 \Delta}{\Delta} \sum_{k=1}^{K} \log \left( \frac{C_2 \Delta^2}{K} \sum_{t=1}^{T} \exp \left( -C_3 \Delta^2 \sum_{s=1}^{t} h_{k,s} \right) \right),
\]

where \( C_1, C_2, \) and \( C_3 \) are positive constants that only depend on \( \sigma \) and \( \hat{\sigma} \).

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, one recovers a lower bound of order $K \Delta \log T$ 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.” 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 that the second arm generates 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:

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

where $\mathbb{E}_\rho$ denotes the expectation with respect to probability measure $\rho$. Using Lemma 2.6 from Tsybakov (2009) 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

$$p_{t,\text{sub}} = \frac{1}{4} \exp \left( -\frac{2\Delta^2}{\sigma^2} \left( \mathbb{E}_{\nu,\nu'}[\tilde{n}_{1,1,t}] + \sum_{s=1}^{t} \frac{\sigma^2}{\hat{\sigma}^2} h_{1,s} \right) \right),$$

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 a mean reward of $-\Delta$, the cumulative regret must also be at least $\Delta \cdot \mathbb{E}_{\nu,\nu'}[\tilde{n}_{1,1,T}]$. Therefore the regret is lower bounded by $\frac{\Delta}{2} \left( \sum_{t=1}^{T} p_{t,\text{sub}} + \mathbb{E}_{\nu,\nu'}[\tilde{n}_{1,1,T}] \right)$, which is greater than $\frac{\Delta^2}{4\Delta} \log \left( \frac{\Delta^2}{2\Delta} \sum_{t=1}^{T} \exp \left( -\frac{2\Delta^2}{\sigma^2} \sum_{s=1}^{t} h_{1,s} \right) \right)$. The argument can be repeated by switching arms 1 and 2. For $K$ arms, we follow the above lines and average over the established bounds to obtain:
\[ R^*_S(H, T) \geq \frac{\sigma^2(K - 1)}{4K\Delta} \sum_{k=1}^{K} \log \left( \frac{\Delta^2}{\sigma^2K} \sum_{t=1}^{T} \exp \left( -\frac{2\Delta^2}{\sigma^2} \sum_{s=1}^{t} h_{k,s} \right) \right), \]

which establishes the result.

3.1 Discussion and subclasses of information flows

Theorem 3.1 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 \( h_{k,t} \)'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 \( \lambda \leq \frac{\hat{\sigma}^2}{4\Delta^2T} \), then
\[
\mathbb{E}_H [R^*_S(H, T)] \geq \frac{\sigma^2(K - 1)}{4\Delta} \log \left( \frac{1 - e^{-1/2}}{\lambda K \sigma^2/\hat{\sigma}^2} \right).
\]

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

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\(^4\). When \( \Delta \) is fixed and independent of the horizon length \( T \), the lower bound scales logarithmically 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).

\(^4\)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).
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.2 and C.3 of the Appendix.

3.1.2 Diminishing information flows

Assume that $h_{k,t}$'s are random variables such that for each arm $k \in K$ and time step $t$,

$$E \left[ \sum_{s=1}^{t} h_{k,s} \right] = \left\lfloor \frac{\sigma^2 \kappa}{2 \Delta^2 \log t} \right\rfloor,$$

for some fixed $\kappa > 0$. 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:

$$R_{\pi}^S(H,T) \geq \frac{\sigma^2(K-1)}{4\Delta} \log \left( \frac{\Delta^2/K\sigma^2}{1 - \kappa} \left( (T+1)^{1-\kappa} - 1 \right) \right).$$

2. If $\kappa > 1$ then:

$$R_{\pi}^S(H,T) \geq \frac{\sigma^2(K-1)}{4\Delta} \log \left( \frac{\Delta^2/K\sigma^2}{\kappa - 1} \left( 1 - \frac{1}{(T+1)^{\kappa-1}} \right) \right).$$

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) 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.4 and C.5 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 decreases over time. The different timing of the $\lambda T$ information arrivals may lead to different regret rates. For example, selecting $\lambda = \frac{\sigma^2 \log T}{\Delta^2 T}$ implies $\kappa = 2$. Then, 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 §4 we show that constant regret is indeed achievable in this setting). This observation conforms 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 is larger), and since earlier information can be used for more decisions (as the remaining horizon is longer)\footnote{This observation can be generalized. The subclasses described in §3.1.1 and §3.1.2 are special cases of the following setting. Let $h_{k,t}$’s be independent random variables such that for each arm $k$ and time $t$, the expected number of information arrivals up to time $t$ satisfies }

The analysis above demonstrates that the achievable performance and optimal policy design depend on the information arrival process: while policies 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 arrival process. In the following sections we address the challenge of adapting to an arbitrary and unknown information arrival processes in the sense of achieving rate-optimal performance without any prior knowledge on the information arrival process.

4 Natural adaptation to the information arrival process

In this section we establish that policies based on posterior sampling and upper confidence bounds may naturally adapt to an a priori unknown information arrival process to achieve the lower bound of Theorem 1 uniformly over the general class of information arrival processes that we consider.

4.1 Robustness of Thompson sampling

Consider the Thompson sampling with Gaussian priors (Agrawal and Goyal 2012, 2013a). In the following adjustment of this policy posteriors are updated both after the policy’s actions as well as after the auxiliary information arrivals. Denote by $n_{k,t}$ and $\bar{X}_{k,n_{k,t}}$ 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:

$$n_{k,t} := \sum_{s=1}^{t-1} 1\{\pi_s = k\} + \sum_{s=1}^{t} \frac{\sigma^2}{2} h_{k,s}, \quad \bar{X}_{k,n_{k,t}} := \frac{\sum_{s=1}^{t} \frac{1}{\sigma^2} 1\{\pi_s = k\} X_{k,s} + \sum_{s=1}^{t} \frac{1}{\hat{\sigma}^2} Z_{k,s}}{\sum_{s=1}^{t} \frac{1}{\sigma^2} 1\{\pi_s = k\} + \sum_{s=1}^{t} \frac{1}{\hat{\sigma}^2} h_{k,s}}. \quad (2)$$

While the expected number of total information arrivals for each arm, $\lambda T$, is determined by the parameter $\lambda$, the concentration of arrivals 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 $\mathbb{E} \left[ \sum_{s=1}^{t} h_{k,s} \right] \sim \lambda T \frac{t^{1-\gamma} - 1}{T^{1-\gamma} - 1}$, corresponding to the class in §3.1.2. Then, when $\lambda T$ is of order $T^{1-\gamma}$ or higher, the lower bound is a constant independent of $T$. 

15
Thompson sampling with auxiliary observations. Inputs: a tuning constant $c$.

1. Initialization: set initial counters $n_{k,0} = 0$ and initial empirical means $\tilde{X}_{k,0} = 0$ for all $k \in K$.
2. At each period $t = 1, \ldots, T$:
   - (a) Observe the vectors $h_t$ and $Z_t$.
   - (b) Sample $\theta_{k,t} \sim N(\tilde{X}_{k,n_{k,t}}, c\sigma^2(n_{k,t} + 1)^{-1})$ for all $k \in K$, and select the arm
     $$\pi_t = \arg \max_k \theta_{k,t}.$$ 
   - (c) Receive and observe a reward $X_{\pi_t,t}$.

The next result establishes that, remarkably, by adjusting the Thompson sampling policy to update posteriors after both the policy’s actions and whenever auxiliary information arrives, one may guarantee it achieves rate optimal performance in the presence of additional information flows.

**Theorem 2 (Near optimality of Thompson sampling with auxiliary observations)** Let $\pi$ be a Thompson sampling with auxiliary observations policy, with $c > 0$. For every $T \geq 1$, and auxiliary information arrival matrix $H$:

$$R_S^\pi(H, T) \leq \sum_{k \in K \setminus \{k^*\}} \left( \frac{C_4}{\Delta_k} \log \left( \sum_{t=0}^T \exp \left( \frac{-\Delta_k^2}{C_4} \sum_{s=1}^t h_{k,s} \right) \right) + \frac{C_5}{\Delta_k^2} + C_6 \Delta_k \right),$$

for some absolute positive constants $C_4$, $C_5$, and $C_6$ that depend on only $c$, $\sigma$ and $\hat{\sigma}$.

**Key ideas in the proof.** To establish the result we decompose the regret associated with each suboptimal arm $k$ into three components: (i) Regret from pulling an arm when its empirical mean deviates from its expectation; (ii) Regret from pulling an arm when its empirical mean does not deviate from its expectation, but $\theta_{k,t}$ deviates from the empirical mean; and (iii) Regret from pulling an arm when its empirical mean does not deviate from its expectation and $\theta_{k,t}$ does not deviate from the empirical mean. Following the analysis in Agrawal and Goyal (2013b) one may use concentration inequalities such as Chernoff-Hoeffding bound to bound the cumulative expressions of cases (i) and (iii) from above by a constant. Case (ii) can be analyzed considering an arrival process with arrival rate that is decreasing exponentially with each arrival. We bound the total number of arrivals of this process (Lemma 11) and establish that this suffices to bound the cumulative regret associated with case (ii).

**Remark 1** 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 $h_{k,t}$’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 for the MAB problem with information flows as a function of the information arrival process:

$$R^*_S(H, T) \leq \sum_{k \in K} \log \left( \sum_{t=0}^{T} \exp \left(-\frac{\Delta_k^2}{2} \frac{t}{C_7} \sum_{s=1}^{t} h_{k,s} \right) \right),$$

where $c$ is a constant that depends on problem parameters such as $K$, $\Delta$, and $\sigma$.

4.2 Robustness of UCB

Consider the UCB1 policy (Auer et al. 2002). In the following adjustment of this policy, mean rewards and counters are updated both after the policy’s actions and after auxiliary information arrivals. Define $n_{k,t}$ and $X_{n_{k,t}}$ as in (2), and consider the next adaptation of UCB1 policy.

**UCB1 with auxiliary observations.** Inputs: a constant $c$.

1. At each period $t = 1, \ldots, T$:
   
   (a) Observe the vectors $h_t$ and $Z_t$
   
   (b) Select the arm
   
   $$\pi_t = \begin{cases} 
   t & \text{if } t \leq K \\
   \arg \max_{k \in K} \left\{ \bar{X}_{k,n_{k,t}} + \sqrt{\frac{\sigma^2 \log t}{n_{k,t}}} \right\} & \text{if } t > K
   \end{cases}$$

   (c) Receive and observe a reward $X_{\pi_t}$

The next result establishes that by adjusting the UCB1 with auxiliary observations policy guarantees rate-optimal performance in the presence of additional information flows.

**Theorem 3** Let $\pi$ be UCB1 with auxiliary observations, tuned by $c > 2$. Then, for any $T \geq 1$, $K \geq 2$, and additional information arrival matrix $H$:

$$R^\pi_S(H, T) \leq \sum_{k \in K} \frac{C_7}{\Delta_k^2} \log \left( \sum_{t=0}^{T} \exp \left(-\frac{\Delta_k^2}{2} \frac{t}{C_7} \sum_{s=1}^{t} h_{k,s} \right) \right) + C_8 \Delta_k,$$

where $C_7$, and $C_8$ are positive constants that depend only on $\sigma$ and $\hat{\sigma}$.

**Key ideas in the proof.** The proof adjusts the analysis of UCB1 in Auer et al. (2002). Pulling a suboptimal arm $k$ at time step $t$ implies that at least one of the following three events occur: (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
\[ \tilde{n}_{k,t-1} \leq \hat{l}_{k,t} - \sum_{s=1}^{t} \frac{\sigma^2}{\Delta^2} h_{k,s}, \]

where \( \hat{l}_{k,t} = \frac{4c\sigma^2 \log(\tau_{k,t})}{\Delta^2} \) with \( \tau_{k,t} := \sum_{s=1}^{t} \exp \left( \frac{\Delta^2}{4c^2 \sigma^2} \sum_{\tau=s}^{t} h_{k,\tau} \right) \), 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:

\[
\sum_{t=1}^{T} \mathbb{1} \left\{ \pi_t = k, \tilde{n}_{k,t-1} \leq \hat{l}_{k,t} - \sum_{s=1}^{t} \frac{\sigma^2}{\Delta^2} h_{k,s} \right\} \leq \max_{1 \leq t \leq T} \left\{ \hat{l}_{k,t} - \sum_{s=1}^{t} \frac{\sigma^2}{\Delta^2} h_{k,s} \right\}. 
\]

Therefore, we establish that for \( c > 2 \),

\[
\mathcal{R}_S^c(\mathbf{H}, T) \leq \sum_{k \in \mathcal{K}\setminus\{k^*\}} \frac{C_7}{\Delta^2_k} \max_{1 \leq t \leq T} \log \left( \sum_{m=1}^{t} \exp \left( \frac{\Delta^2_k}{C_7} \sum_{s=m}^{t} h_{k,s} - \frac{\Delta^2_k}{C_7} \sum_{s=1}^{t} h_{k,s} \right) \right) + C_8 \Delta_k. 
\]

### 4.3 Empirical proof of concept using content recommendations data

To demonstrate the value that may be captured in practice by leveraging auxiliary information flows, we use data from a large US media site to empirically evaluate the performance that can be achieved when recommending articles that have unknown impact on the future browsing path of readers. Two key performance indicators in this setting are the likelihood of a reader to click on a recommendation and, when clicking on it, the likelihood of the reader to continue using the recommendation service by clicking on another recommendation (Besbes et al. 2016). The operational challenge in managing content recommendations is therefore dual. A first challenge is to estimate these indicators early on when new content is introduced. But another key challenge, on which we focus here, is that the likelihood of readers to continue consuming content after reading an article is driven by article features (e.g., length, number of photos) that may change after initial estimates for these indicators have been established.

The objective of the following analysis is to show how the performance of sequential experimentation, after a change in article features, can improve by utilizing observations of actions taken by readers that arrived to that article from external search. Our data set includes a list of articles, and consists of:

(i) times in which these article were recommended and the in-site browsing path that followed these recommendations; and

(ii) times at which readers arrived to these articles directly from external search engines (such as Google) and the browsing path that followed these visits.

**Setup.** While we defer the complete setup description to Appendix E, we next provide its key elements. We followed a one-armed bandit setting for analyzing sequential experimentation with each article in the presence of a known outside option. We adopted the times the article was recommended from first position (highest in a list of 5 recommended links) as the sequence of decision epochs for experimenting
with that article. (We focused on the 200 articles that were most frequently recommended in a given
day, and for consistency set $T = 2,000$ for each, and ignored subsequent epochs.)

For a fixed article $a$ and day $d$, denote by $\text{CTR}_{a,d}$ the fraction of readers that clicked the article’s title
during day $d$ when this article was recommended to them in first position (click-through rate), and by
$\text{CVR}_{a,d}^{\text{recom}}$ the fraction of readers that clicked on a recommendation hosted by the article during day $d$
after arriving to that article through content recommendation (convergence rate).

When arm $k \in \{0, 1\}$ is pulled at time $t$, the agent observes a reward equal to the (revenue-
normalized) one-step lookahead recommendation value, $X_{k,t}^{(1)} \left(1 + X_{k,t}^{(2)}\right)$, where $X_{k,t}^{(1)} \sim \text{Ber}(\text{CTR}_{a,d,k})$ and $X_{k,t}^{(2)} \sim \text{Ber}(\text{CVR}_{a,d,k}^{\text{recom}})$. The unknown arm has a conversion rate equal to $\text{CVR}_{a,d}^{\text{recom}}$, and the
outside option has a known conversion rate that is slightly greater or smaller than $\text{CVR}_{a,d}^{\text{recom}}$. The
baseline conversion rate $\text{CVR}_{a,d}^{\text{recom}}$ was calculated from the data, but is unknown to the recommender
system. We assume that click-through rates are already known; for simplicity we assume that they are
fixed and equal across both arms, but note that results are robust with respect to this assumption.

While the observations above are drawn randomly, the matrix $H$ and sample path of auxiliary
observations are $\text{fixed and determined from the data}$ as follows. We extracted the trajectory of information
arrivals $\{h_{1,t}\}$ from the number of readers that arrived to the article from an external search engine
between consecutive decision epochs. Information flows in the data often include two or more auxiliary
observations between consecutive decision epochs; as we comment in §2.2, the performance bounds we
establish in this paper indeed hold for any integer values assigned to entries of the matrix $H$. For each
epoch $t$ and arrival-from-search $m \in \{1, \ldots, h_{1,t}\}$ we denote by $Y_{1,t,m} \in \{0, 1\}$ an indicator of whether
the reader clicked on a recommendation hosted by the article. We denote by $\text{CVR}_{a,d}^{\text{search}}$ the fraction of
readers that clicked on a recommendation hosted by article $a$ during day $d$ after arriving to that article
from search, and define:

$$\alpha_{a,d} := \frac{\text{CVR}_{a,d}^{\text{recom}}}{\text{CVR}_{a,d}^{\text{search}}}.$$  

When the CTR is the same across alternatives, one may look only on conversions feedback with
$Z_{1,t,m} = \alpha_{a,d} Y_{k,t,m}$ (see Appendix E for details). Then, the auxiliary observations $\{Y_{1,t,m}\}$ belong to the
class of linear mappings from §2.1 with $\beta_1 = 0$, and $\alpha_1 = \alpha_{a,d}$.

We construct an estimator of $\alpha_{a,d}$ based on $\alpha_{a,d-1}$, the fraction of the two conversion rates from the
previous day. Note that, as a fraction, $\alpha_{a,d-1}$ is not an unbiased estimator of $\alpha_{a,d}$. We assume that
$(\alpha_{a,d-1}/\alpha_{a,d})$ is a log-normal random variable, that is, $\alpha_{a,d-1} = \alpha_{a,d} \cdot \exp\{\tilde{\sigma}^2 W\}$ for $W \sim \mathcal{N}(0, 1)$ and some $\tilde{\sigma} > 0$, and construct the following unbiased estimator of $\alpha_{a,d}$:

6This one-step lookahead payoff was identified in Besbes et al. (2016) as a practical approximation for the full lookahead
value generated by a content recommendation.

7Conversion rates of readers that clicked on a recommendation are typically higher than conversion rates of readers that
arrived to the content page directly from search; see a related discussion in Besbes et al. (2016) on “experienced” versus
“inexperienced” readers. In our data set, values of $\alpha_{a,d}$ are in the range $[1, 16]$.  

19
\[ \hat{\alpha}_{a,d} = \alpha_{a,d-1} \cdot \exp \left\{ -\hat{\sigma}^2 / 2 \right\}. \]

**Results.** We compare a standard Thompson Sampling to the variant from §4.2 that utilizes auxiliary observations based on the estimates \( \hat{\alpha} \). The right hand side of Figure 2 identifies a log-normal distribution of the fraction \( \hat{\alpha}/\alpha \), with \( \hat{\sigma}^2 \approx 0.53 \). This implies that practical estimators for the mapping \( \phi \) could be contracted based on the previous day, but with non-trivial estimation errors. We note that these errors could be significantly reduced using additional data and more sophisticated estimation methods.

The performance comparison appears on the left side of Figure 2. Each point on the scatter plot corresponds to the normalized long-run average regret for a certain article on a certain day. The performance of the standard Thompson sampling is independent of additional information arrivals. The performance of the version that utilizes auxiliary information is comparable to the standard version with very little auxiliary observations, and significantly improves when the amount of auxiliary observations increases. Main sources of variability in performance are (i) variability in the baseline conversion rates \( \text{CVR}_{a,d} \), affecting both algorithms; and (ii) variability in the estimation quality of \( \alpha \), affecting only the version that uses the auxiliary observations. Utilizing information flows led to regret reduction of 27.8% on average; We note that this performance improvement is achieved despite (i) suffering occasionally from inaccuracies in the estimate of the mapping \( \phi(\cdot) \), and (ii) tuning both algorithms by the value \( c \) that optimizes the performance of the standard Thompson sampling.

![Figure 2: (Left) Performance of a standard Thompson Sampling that ignores auxiliary observations relative to one that utilizes auxiliary observations, both tuned by \( c = 0.25 \). Each point corresponds to the (normalized) mean cumulative regret averaged over 200 replications for a certain article on a certain day, and the fixed sample path of auxiliary observations for that article on that day. (Right) Histogram of \( \log(\alpha_{a,d-1}/\alpha_{a,d}) \), with the fitted normal distribution suggesting \( \alpha_{a,d-1} \sim \text{Lognormal}(\log \alpha_{a,d}, \hat{\sigma}^2) \) with \( \hat{\sigma}^2 \approx 0.53 \).](image)

**4.4 Discussion**

Theorems 2 and 3 establish that Thompson sampling and UCB polices guarantee the best achievable performance (up to some multiplicative constant) under any arbitrary information arrival process captured by the matrix \( H \); expressions obtained by Thompson sampling for the settings described in
that match the lower bounds established in §3.1.1 and §3.1.2 for any values of λ and κ are given in Appendix C.1 (see Corollaries 1 and 2). These results imply that both policies exhibit remarkable robustness with respect to the information collection process, and significantly generalize the class of information collection processes to which these policies have been applied thus far. Nevertheless, other policies such as ε_t-greedy-type policies, do not exhibit such robustness (as demonstrated in §5.1).

From a practical standpoint this is a caveat, as due to considerations such as computational complexity and latency, simple algorithms akin to ε_t-greedy may be preferable for large-scale online recommender systems; see, e.g., Basilico (2013) for a discussion on the necessity of algorithmic simplicity for large-scale recommender systems, Kohli et al. (2013) and Tang et al. (2015) for performance analysis of ε_t-greedy in various product recommendation settings, and the policy run-time analysis in Agarwal et al. (2014).

Another advantage of ε_t-greedy-type policies is that they allow the collection of independent samples (without inner-sample correlation) and by that facilitate statistical inference, whereas other policies such as Thompson sampling and UCB introduce bias when “standard” inference methods are applied; see, e.g., the analysis in Nie et al. (2017) and the empirical study in Villar et al. (2015).

We therefore devote §5 to developing a general adaptive method for adjusting the experimentation schedule of ε_t-greedy-type policies in the presence of additional information flows. Our approach can be further applied to other MAB frameworks, including ones that are well-grounded in the practice of product recommendations. We demonstrate this in §6 through considering an extension of the contextual bandits model from Goldenshluger and Zeevi (2013), and establishing improved performance guarantees in the presence of auxiliary information flows.

5 Adaptive ε_t-greedy policy via virtual time indexes

5.1 Policy design intuition

In this section we first demonstrate, through the case of ε_t-greedy policy, that classical policy design may in fact fail to achieve the lower bound in Theorem 1 in the presence of information flows, and develop intuition for how this policy may be adjusted to better leverage auxiliary information through the virtual time indexes method. Formal policy design and analysis will follow in §5.2.

The inefficiency of a naive adaptation of ε_t-greedy. Despite the robustness that was established in §4 for policies that are based on posterior sampling or upper confidence bounds, in general, the approach of accounting for auxiliary information while otherwise maintaining the policy structure may not suffice for achieving the lower bound established in Theorem 1. To demonstrate this, consider the ε_t-greedy policy (Auer et al. 2002), which at each period t selects an arm randomly with probability ε_t that is proportional to t^{-1}, and with probability 1 − ε_t selects the arm with the highest reward estimate.
Without auxiliary information, \( \epsilon_t \)-greedy guarantees rate-optimal regret of order \( \log T \), but this optimality does not naturally carry over to settings with auxiliary information. Consider, for example, the class of stationary information flows described in §3.1.1 with an arrival rate \( \lambda \geq \frac{\hat{\sigma}^2}{4 \Delta^2 T} \). While constant regret is achievable under this class, \( \epsilon_t \)-greedy explores with suboptimal arms at a rate that is independent of the number of auxiliary observations, and therefore still incurs regret of order \( \log T \).

**Over-exploration in the presence of additional information.** For simplicity, consider a 2-armed bandit problem with \( \mu_1 > \mu_2 \). At each time \( t \) the \( \epsilon_t \)-greedy policy explores over arm 2 independently with probability \( \epsilon_t = ct^{-1} \) for some constant \( c > 0 \). As a continuous-time proxy for the minimal number of times arm 2 is selected by time \( t \), consider the function
\[
\int_{s=1}^{t} \epsilon_s ds = c \int_{s=1}^{t} \frac{ds}{s} = c \log t.
\]
The probability of best-arm misidentification at period \( t \) can be bounded from above by
\[
\exp \left( -\bar{c} \int_{s=1}^{t} \epsilon_s ds \right) \leq \tilde{c} t^{-1},
\]
for some constants \( \bar{c}, \tilde{c} \). Thus, the selection \( \epsilon_t = ct^{-1} \) balances losses due to exploration and exploitation.

Next, assume that just before time \( t_0 \), an additional independent reward observation of arm 2 is collected. Then, at time \( t_0 \), the minimal number of observations from arm 2 increases to \( (1 + c \log t_0) \), and the upper bound on the probability of best-arm misidentification decreases by factor \( e^{-\tilde{c}} \):
\[
\exp \left( -\bar{c} \left( 1 + \int_{s=1}^{t_0} \epsilon_s ds \right) \right) \leq e^{-\tilde{c}} \cdot \tilde{c} t^{-1}, \quad \forall t \geq t_0.
\]
Therefore, when there are many auxiliary observations, the loss from best-arm misidentification is guaranteed to diminish, but regret of order \( \log T \) is still incurred due to over-exploration.

**Endogenizing the exploration rate.** Note that there exists a future time period \( \hat{t}_0 \geq t_0 \) such that:
\[
1 + c \int_{1}^{t_0} \frac{ds}{s} = c \int_{1}^{\hat{t}_0} \frac{ds}{s}.
\]
In words, the minimal number of observations from arm 2 by time \( t_0 \), including the one that arrived just before \( t_0 \), equals (in expectation) the minimal number of observations from arm 2 by time \( \hat{t}_0 \) without any additional information arrivals. Therefore, replacing \( \epsilon_t = ct^{-1} \) with \( \epsilon_t = c\hat{t}^{-1} \) for all \( t \geq t_0 \) would adjust the exploration rate to fit the amount of information actually collected. Exploration and exploitation

---

8 This upper bound on the probability of misidentifying the best arm can be obtained using standard concentration inequalities and is formalized, for example, in Step 5 of the proof of Theorem 4.

9 The principle of designing the exploration schedule to balance losses from experimenting with suboptimal arms and from misidentification of the best arm is commonly used for developing efficient policies; see, e.g., related discussions in Auer et al. (2002), Langford and Zhang (2008), Goldenshluger and Zeevi (2013), and Bastani and Bayati (2015).
would be balanced once again, with reduced regret; see Figure 4 for an illustration of this observation. We therefore adjust the exploration rate to be $\epsilon_t = c(\tau(t))^{-1}$ for some virtual time $\tau(t)$. We set $\tau(t) = t$ for all $t < t_0$, and $\tau(t) = \hat{t}_0 + (t - t_0)$ for $t \geq t_0$. Solving (3) for $\hat{t}_0$, we write $\tau(t)$ in closed form:

$$\tau(t) = \begin{cases} 
    t & t < t_0 \\
    c_0 t_0 + (t - t_0) & t \geq t_0,
\end{cases}$$

for some constant $c_0 > 1$. Therefore, the virtual time grows together with $t$, and advanced by a multiplicative constant whenever an auxiliary observation is collected.

### 5.2 A rate-optimal adaptive adjustment of $\epsilon_t$-greedy

We apply the ideas discussed in §5.1 to design an $\epsilon_t$-greedy policy with adaptive exploration that dynamically adjusts the exploration rate in the presence of unknown information flows. Define $n_{k,t}$ and $X_{n_{k,t}}$ as in (2), and consider the following adaptation of the $\epsilon_t$-greedy policy.

**$\epsilon_t$-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 $h_t$ and $Z_t$, and update virtual time indexes for all $k \in K$:

   $$\tau_{k,t} = (\tau_{k,t-1} + 1) \cdot \exp \left( \frac{h_{k,t} \Delta^2}{c\sigma^2} \right)$$
Figure 4: Illustration of the adaptive exploration approach. (Left) Virtual time index $\tau(\cdot)$ is advanced using multiplicative factors whenever auxiliary information arrives. (Right) Exploration rate decreases as a function of $\tau(\cdot)$, exhibiting discrete “jumps” whenever auxiliary information is collected.

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

$$\pi_t = k \quad \text{with probability} \quad \frac{\tau_{k,t}^{-1}}{\sum_{k'=1}^{K} \tau_{k',t}^{-1}}, \quad \text{for all } k \in \mathcal{K}$$

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

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

(c) Receive and observe a reward $X_{\pi_t,t}$

At every period $t$ the $\epsilon_t$-greedy with adaptive exploration policy dynamically reacts to the information sample path by advancing virtual time indexes associated with different arms based on auxiliary observations that were collected since the last period. Then, the policy explores with probability that is proportional to $\sum_{k'=1}^{K} \tau_{k',t}^{-1}$, and otherwise pulls the arm with the highest empirical mean reward.

Every time additional information on arm $k$ is observed, a carefully selected multiplicative factor is used to advance the virtual time index $\tau_{k,t}$ according to the update rule $\tau_{k,t} = (\tau_{k,t-1} + 1) \cdot \exp(\delta \cdot h_{k,t})$, for some suitably selected $\delta$. In doing so, the policy effectively reduce exploration rates in order to explores over each arm $k$ at a rate that would have been appropriate without auxiliary information flows at a future time step $\tau_{k,t}$. This guarantees that the loss due to exploration is balanced with the loss due to best-arm misidentification throughout the horizon. Advancing virtual times based on the information sample path and the impact on the exploration rate of a policy are illustrated in Figure 4.

The following result characterizes the guaranteed performance and establishes the rate optimality of $\epsilon_t$-greedy with adaptive exploration in the presence of unknown information flows.
Theorem 4 (Near optimality of $\epsilon_t$-greedy with adaptive exploration) Let $\pi$ be an $\epsilon_t$-greedy with adaptive exploration policy, tuned by $c > \max \left\{ 16, \frac{10\Delta^2}{\sigma^2} \right\}$. Then, for every $T \geq 1$, and for any information arrival matrix $H$, one has:

$$R_\pi^H(H, T) \leq \sum_{k \in K} \Delta_k \left( C_9 \frac{\Delta^2}{\sigma^2} \log \left( \sum_{t=t^*+1}^T \exp \left( -\frac{\Delta^2}{C_9} \sum_{s=1}^{t} h_{k,\tau} \right) \right) + C_{10} \right),$$

where $C_9$, and $C_{10}$ are positive constants that depend only on $\sigma$ and $\hat{\sigma}$.

Key ideas in the proof. The proof decomposes regret into exploration and exploitation time periods. To bound the regret at exploration time periods express virtual time indexes as

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

Denoting by $t_m$ the time step at which the $m^{th}$ auxiliary observation for arm $k$ was collected, 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 $\frac{c\sigma^2}{\Delta^2} \log \left( \frac{2\Delta}{\Delta_k} \sum_{k=1}^{t_{m+1}} h_{k,m} \right) - 1$. Summing over all values of $m$, we obtain that the regret over exploration time periods is bounded from above by

$$\sum_{k \in K} \Delta_k \cdot \frac{c\sigma^2}{\Delta^2} \log \left( 2 \sum_{t=0}^T \exp \left( -\frac{\Delta^2}{c\sigma^2} \sum_{s=1}^{t} h_{k,\tau} \right) \right).$$

To analyze regret at 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. When $c > \max \left\{ 16, \frac{10\Delta^2}{\sigma^2} \right\}$, this regret component decays at rate of at most $t^{-1}$.

5.3 Numerical analysis

We further evaluate the performance of policies detailed in §4 and §5 through numerical experiments that are detailed in Appendix D. First, we compared the performance of the adjusted $\epsilon_t$-greedy described earlier with a more “naive” version of this policy that updates the empirical means and the reward observation counters upon auxiliary information arrivals but do not update the time index as prescribed, and found that using virtual time indexes always empirically outperforms using standard time index.

We compared the performance of UCB with a standard time index and the one of a UCB policy that uses dynamic virtual time indexes; to that extent in Theorem 6 (Appendix D) we first establish the rate optimality of UCB with virtual time indexes. We observe that with a lot of auxiliary information, UCB with standard time index may incur less regret, whereas with little auxiliary information UCB with virtual time indexes performs better. The reason for this phenomenon is that virtual time indexes shrink
the upper confidence bound slower relative to a standard time index, which drives more exploration.

We explored how the selection of the tuning parameter $c$ impacts the performance of policies. For all the policies discussed in this paper, we found that when the auxiliary information is abundant, a lower tuning parameter $c$ typically exhibits better performance. However, when there is less auxiliary information, a very small tuning parameter $c$ may result in a relatively high regret.

We also analyze the regret incurred by policies over shorter time horizons, tested for robustness with respect to misspecification of the gap parameter $\Delta$, and evaluate an approach for estimating $\Delta$ throughout the decision process based on auxiliary observations, which we discuss in §7.2.

6 Application to contextual MAB

In this section we extend our model and method to the contextual MAB framework with linear response and stochastic contexts that is described in [Goldenshluger and Zeevi (2013)].

6.1 Linear response model with additional information

Following the formulation in [Goldenshluger and Zeevi (2013)], the analysis in this section focuses on the case of two arms, but can be directly extended to any finite number of arms. Assume that at each time step $t$, the agent first observes a random context $W_t \in \mathbb{R}^{d+1}$ such that the first element of $W_t$ is always 1 (namely, $(W_t)_1 = 1$) and then selects an arm. If arm $k$ is selected, the reward $X_{k,t}$ is observed and received, where $X_{k,t}$ is assumed to be $\sigma^2$-sub-Gaussian, and to have mean $E[X_{k,t}] = W_t^\top \beta_k$, where $\beta_k$ is a vector of unknown parameters. Rewards are assumed to be independent across time steps and arms.

As defined in §2, before each time step $t$, the agent may or may not observe auxiliary information on some of the arms. Let $h_{k,t} \in \{0,1\}$ denote the indicator of observing auxiliary information on arm $k$ before time $t$. If $h_{k,t} = 1$ then, a random context $V_{k,t} \in \mathbb{R}^{d+1}$ is observed, such that the first element of $V_{k,t}$ is always 1 (that is, $(V_{k,t})_1 = 1$), along with an outcome $Y_{k,t}$, which is $\hat{\sigma}^2$-sub-Gaussian and has mean $E[Y_{k,t}] = V_{k,t}^\top \beta_k$. The outcomes $Y_{k,t}$ are assumed to be independent across time epochs and arms.

An admissible policy $\pi$ is a sequence of mappings $\{\pi_t : t = 1, 2, \ldots\}$ from past observations and current contexts to arms. We measure performance relative to a dynamic benchmark that selects the best arm at each epoch $t$ given the context $W_t$ and prior knowledge of the vectors $\beta_k$, $k \in K$, and denote it by $\pi^*_t := \arg \max_k W_t^\top \beta_k$. For a given arrival matrix $H$, this regret measure is defined as follows:

---

10While studies such as [Li et al. (2010)] and [Agrawal and Goyal (2013b)] analyze a contextual MAB problem with adversarial contexts using upper-confidence bound or posterior sampling techniques, to the best of our knowledge, these techniques have not been deployed for addressing problems with stochastic contexts.

11Formally, $\pi_t$ is a mapping from $\mathcal{H}_t := \left\{ \begin{array}{ll}
\{W_1, ((h_{k,1}, h_{k,1} Y_{k,1}, h_{k,1} V_{k,1})_{k \in K}) \} & \text{if } t = 1 \\
\{X_{s,t} \}_{s=1}^{t-1}, \{W_s\}_{s=1}^{t-1}, \{((h_{k,s}, h_{k,s} Y_{k,s}, h_{k,s} V_{k,s}))_{s \leq t, k \in K} \} & \text{if } t > 1
\end{array} \right\}$ to $K$. 

\[ R^\pi (H, T) := \mathbb{E}^\pi \left[ \sum_{t=1}^T X_{\pi^*_t, t} - X_{\pi_t, t} \right]. \]

We assume that the contexts \( W_t \) are generated independently according to the distribution \( P_w \), and that for each \( k \) the contexts \( V_{k,t} \) are generated independently according to the distribution \( P_{v}^{(k)} \). Following Goldenshluger and Zeevi (2013), we further make the following assumptions.

**Assumption 1** There exists a positive \( r \) such that \( \max_{i \geq 2} |(W_t)_i| \leq r \). Furthermore, there exists a positive \( \lambda \) such that \( \lambda_{\min}(Q) \geq \lambda > 0 \), where \( Q := \mathbb{E} [W_t W_t^\top] \).

**Assumption 2** There exists positive constants \( \rho_0, L \) and \( \alpha \) such that \( \mathbb{P} \{|(\beta_1 - \beta_2)^\top W_t| \leq \rho\} \leq L \rho^\alpha \), for all \( \rho \in (0, \rho_0] \).

Assumption 2, known as the margin condition in the statistics literature (see, e.g., Tsybakov 2004), determines how fast the covariate mass around the decision boundary shrinks. When \( \alpha = 1 \) this mass shrinks linearly, when \( \alpha > 1 \) it shrinks super-linearly, and when \( \alpha < 1 \) it shrinks sub-linearly. The larger the parameter \( \alpha \) the easier the problem is since it becomes less likely to have a covariate near the decision boundary where it is hard to identify the optimal arm.

**Assumption 3** There exists a positive constant \( \lambda_{\star} \) such that \( \lambda_{\min} (\mathbb{E} U_{t}^+ W_t W_t^\top) \land \lambda_{\min} (\mathbb{E} U_{t}^- W_t W_t^\top) \geq \lambda_{\star} > 0 \), where \( U_{t}^+ = 1 \{(\beta_1 - \beta_2)^\top W_t \geq \rho_0\} \) and \( U_{t}^- = 1 \{(\beta_2 - \beta_1)^\top W_t \geq \rho_0\} \).

**Assumption 4** There exists some positive constant \( b \) such that \( \|\beta_k\|_{\infty} \leq b \) for all \( k \in \mathcal{K} \).

In addition, we make the following assumption on the covariates corresponding to the additional information, which corresponds to Assumption 1.

**Assumption 5** \( \max_{i \geq 2} |(V_{k,t})_i| \leq r \) for the same constant \( r \) in Assumption 1. Furthermore, \( \lambda_{\min}(\hat{Q}_k) \geq \lambda > 0 \), where \( \hat{Q}_k := \mathbb{E} [V_{k,t} V_{k,t}^\top] \) for the same constant \( \lambda \) in Assumption 1.

Under assumptions 1-4, Goldenshluger and Zeevi (2013) propose a policy based on ordinary least squares that achieves logarithmic regret. They also show that when Assumption 2 holds with \( \alpha = 1 \), the logarithmic regret they achieve is optimal. In fact, Goldenshluger and Zeevi (2013) show that when \( \alpha = 1 \) even with full feedback (as opposed to bandit feedback) no policy can guarantee a better regret rate. While the policy suggested in Goldenshluger and Zeevi (2013) guarantees logarithmic regret also when \( \alpha > 1 \), they suggest that this regret rate can be improved. In the remainder of this

---

12 This conjecture is based on prior analysis from Goldenshluger and Zeevi (2009) that consider a less general one-armed bandit framework with a single covariate and show that when it is known that Assumption 2 holds with \( \alpha > 1 \) constant regret is achievable using a semi-myopic policy. Nevertheless, their semi-myopic policy is sub-optimal if Assumption 2 holds with \( \alpha = 1 \). To the best of our knowledge, even for the one-armed bandit setting in Goldenshluger and Zeevi (2009) no single policy was shown to achieve both logarithmic regret when Assumption 2 holds with \( \alpha = 1 \), as well as improved regret rate when Assumption 2 holds with \( \alpha > 1 \) (without prior knowledge on the value of \( \alpha \) for which Assumption 2 holds).
section we adjust the sampling method from Goldenshluger and Zeevi (2013) following the adaptive exploration approach proposed in this paper. When Assumption 2 holds with \( \alpha = 1 \) the adjusted policy achieves logarithmic regret, but when Assumption 2 holds with \( \alpha > 1 \) the regret it guarantees reduces to \( O \left( \sum_{k \in K} \log \left( \frac{T}{\exp \left( -c \cdot \sum_{s=1}^{t} h_{k,s} \right)} \right) \right) \) for some constant \( c \) and, in particular, decreases to a constant when auxiliary information appears frequently and early enough.\(^{13}\)

We note that in this section we do not assume any mapping \( \phi_k \) for extracting information from outcomes \( Y_{k,t} \). Rather, this information is used for enhancing estimates of the underlying parameters and improving performance through leveraging the parametric structure of the contextual MAB problem.

### 6.2 Adjusted policy and adaptive performance guarantees

We first introduce additional notation that is required for describing the proposed policy. Let \( q \) and \( \hat{q} \) be some positive constants. Define \( t^* := \max \left\{ \frac{2K}{q^2}, K(d + 1) \right\} \), and define the counter \( \hat{n}_{k,t} \) as follows:

\[
\hat{n}_{k,t} := \sum_{s=1}^{t} \mathbb{1} \left\{ \pi_s = k, \ G_t \leq \frac{2}{q} \sum_{k'=1}^{K} \tau_{k',t}^{-1} \right\} + \sum_{s=1}^{t+1} \frac{\hat{q}}{q} h_{k,s},
\]

where \( \tau_{k,t} \) is the virtual time index, and let \( n_{k,t} \) be the number of times arm \( k \) is observed up to time \( t \):

\[
n_{k,t} := \sum_{s=1}^{t} \mathbb{1} \{ \pi_s = k \} + \sum_{s=1}^{t+1} \frac{\hat{q}}{q} h_{k,s}.
\]

Using these two counters, we further define:

\[
\hat{Q}_{k,t} := \left( \sum_{s=1}^{t-1} \mathbb{1} \left\{ \pi_s = k, \ G_t \leq \frac{2}{q} \sum_{k'=1}^{K} \tau_{k',t}^{-1} \right\} \cdot W_s W_s^\top + \sum_{s=1}^{t} \frac{\hat{q}}{q} h_{k,s} \cdot V_{k,s} V_{k,s}^\top \right) / \hat{n}_{k,t-1},
\]

\[
\hat{\beta}_{k,t} := \hat{Q}_{k,t}^{-1} \cdot \frac{1}{\hat{n}_{k,t-1}} \left[ \sum_{s=1}^{t-1} \mathbb{1} \left\{ \pi_s = k \right\} \cdot W_s W_s^\top + \sum_{s=1}^{t} \frac{\hat{q}}{q} h_{k,s} \cdot V_{k,s} V_{k,s}^\top \right],
\]

\[
\hat{Q}_{k,t} := \left( \sum_{s=1}^{t-1} \mathbb{1} \{ \pi_s = k \} \cdot W_s W_s^\top + \sum_{s=1}^{t} \frac{\hat{q}}{q} h_{k,s} \cdot V_{k,s} V_{k,s}^\top \right) / n_{k,t-1},
\]

\[
\hat{\beta}_{k,t} := \hat{Q}_{k,t}^{-1} \cdot \frac{1}{n_{k,t-1}} \left[ \sum_{s=1}^{t-1} \mathbb{1} \{ \pi_s = k \} \cdot W_s X_{k,s} + \sum_{s=1}^{t} \frac{\hat{q}}{q} h_{k,s} \cdot V_{k,s} Y_{k,s} \right].
\]

**OLS bandit with \( \varepsilon_t \)-greedy adaptive exploration.** Input: the tuning parameters \( h, q, \) and \( \hat{q} \).

**Initialization:** \( \tau_{k,0} = 0 \) and \( \hat{n}_{k,0} = 0 \) for all \( k \in \mathcal{K} \).

At each period \( t = 1, 2, \ldots, T \):

\(^{13}\)We note that Bastani et al. (2017) show that when Assumption 3 is replaced with a stronger assumption on the diversity of contexts, even a myopic policy is guaranteed to achieve rate optimal performance both when \( \alpha > 1 \) and \( \alpha = 1 \), achieving constant regret in the former case and logarithmic regret in the latter case.
Observe the vectors $h_t$ and $h_k,tV_{k,t}$ for $k \in K$ and update the virtual time indexes for all $k \in K$:

\[
\tau_{k,t} = \begin{cases} 
\tau_{k,t-1} + 1 & \text{if } t < t^* \\
(\tau_{k,t-1} + 1) \cdot \exp \left( \frac{t}{2} \sum_{s=1}^t \frac{2h_{k,s}}{q} \right) & \text{if } t = t^* \\
(\tau_{k,t-1} + 1) \cdot \exp \left( \frac{2h_{k,t}}{q} \right) & \text{if } t > t^*
\end{cases}
\]

(b) Observe context $W_t$, and generate a random variable $G_t \sim \text{Uniform}[0,1]$

(c) If $t \leq t^*$ alternate over arms using $\pi_t = t \mod K$. Otherwise, if $G_t \leq \frac{2}{q} \sum_{k'=1}^K \tau_{k',t}^{-1}$ select an arm at random:

\[
\pi_t = k \quad \text{with probability } \frac{\tau_{k,t}^{-1}}{\sum_{k'=1}^K \tau_{k',t}^{-1}}, \quad \text{for all } k \in K.
\]

Otherwise, if $|(\hat{\beta}_{1,t} - \hat{\beta}_{2,t})^T W_t| > h/2$ set

\[
\pi_t = \arg \max_{k \in K} \hat{\beta}_{k,t}^T W_t.
\]

Otherwise, set

\[
\pi_t = \arg \max_{k \in K} \tilde{\beta}_{k,t}^T W_t.
\]

Similar to the $\epsilon_t$-greedy with adaptive exploration policy, the adjusted OLS bandit policy dynamically reacts to the arrival of additional information to balance between exploration and exploitation. Our policy adjust the fixed sampling schedule suggested in [Goldenshluger and Zeevi (2013)], to make this schedule adaptive with respect to the information arrival: whenever auxiliary information arrives, it advances the exploration schedule as if the policy itself has collected all the samples.

The following result establishes that OLS bandit with adaptive exploration achieves logarithmic regret when Assumption 2 holds with $\alpha = 1$, but also, it achieves improved regret (in particular, constant regret, when information arrives frequently/early enough) when Assumption 2 holds with $\alpha > 1$, without prior knowledge on the value of $\alpha$ for which Assumption 2 holds.

**Theorem 5** Let Assumptions 2 hold. Fix $T \geq 1$ and an information arrival matrix $H$, and let $\pi$ be an OLS bandit with $\epsilon_t$-greedy adaptive exploration policy, tuned by the parameters $h \leq \rho_0$, as well as

\[
q \leq \min \left\{ \frac{h^2 \lambda}{192 \sigma^2 (1 + r^2 d)}, \frac{\lambda^2}{24 (d + 1)^2 \max\{r^2, r^4\}}, \frac{1}{5} \right\}; \quad \hat{q} \leq \min \left\{ \frac{h^2 \lambda}{192 \sigma^2 (1 + r^2 d)}, \frac{\lambda^2}{24 (d + 1)^2 \max\{r^2, r^4\}} \right\}.
\]

If Assumption 3 holds with $\alpha = 1$, then

\[
\mathcal{R}^\pi (H, T) \leq \frac{4b}{q} \sqrt{\lambda_{\max}(Q)} \log(T + 1) + C_{11} Lo^2 d^2 (1 + rd^2) \lambda^2 \max\{1, r^2\} \log T + C_{12}.
\]
If Assumption holds with $\alpha > 1$, then

$$R^\pi (H, T) \leq \sum_{k \in K} \frac{2b}{q} \sqrt{\lambda_{\text{max}}(Q)} \log \left( \sum_{t=0}^{T} \exp \left( -\frac{q^2}{2} \sum_{s=1}^{t} h_{k,s} \right) \right) + C_{13},$$

where $C_{11}$ is an absolute constant, and $C_{12}, C_{13}$ are positive constants that depend on $\sigma^2, b, r, \alpha$ and $d$.

## 7 Concluding remarks and extensions

In this paper we considered an extension of the multi-armed bandit framework, allowing for unknown and arbitrary information flows. We studied the impact of auxiliary information on the design of efficient learning policies and on the performance that can be achieved. Through matching lower and upper bounds we identified the minimax (regret) complexity of this class of problems, which provides a sharp criterion for identifying rate-optimal policies. We established that Thompson sampling and UCB1 policies can be leveraged to achieve rate-optimality and, in that sense, exhibit remarkable robustness with respect to the information arrival process. We further introduced a virtual time indexes method to adjust the $\varepsilon_t$-greedy policy by endogenously controlling its exploration rate in a manner that guarantees rate-optimality, and showed how the virtual time indexes method can be used in the more general setting of contextual bandits with linear response. In addition, using content recommendations data, we demonstrated the value that can be captured in practice by leveraging auxiliary information flows.

We next discuss some additional extensions. In §7.1 we consider the case where the mappings $\phi_k$, connecting auxiliary observations and mean rewards, are a priori unknown. We provide an impossibility result, showing that in general, when these mappings are unknown, no policy can guarantee any performance improvement by using auxiliary information flows. In §7.2 we demonstrate how auxiliary information flows could be used for estimating the gaps $\Delta_k$ throughout the sequential decision process. In §7.3 we establish that our approach leads to improved performance bounds also in a broad class of information flows that depend on the history. In §7.4 we conclude with some open questions.

### 7.1 Unknown mappings $\phi_k$

Our basic formulation considers known mappings $\phi_k$ such that $\mathbb{E} [\phi_k(Y_{k,t})] = \mu_k$ for all $k \in K$. One interesting question is what type of performance guarantees could be established when the mappings $\phi_k$ are a priori unknown. For that matter we state an impossibility result in terms of the performance gain one could hope to extract from additional information when these mappings are unknown. We establish that for any information arrival process $H$, the regret can grow logarithmically with the horizon length $T$. In other words, no admissible policy can guarantee any performance gain based on additional information when the mappings $\phi_k$ that connect this information to the mean rewards are a priori unknown.
Proposition 1 (No guaranteed performance gain with unknown mapping \(\phi_k\)) Assume that the mappings \(\{\phi_k; k \in \mathcal{K}\}\) are unknown. For any \(T \geq 1\) and information arrival matrix \(H\), the worst-case regret of any admissible policy \(\pi \in \mathcal{P}\) is bounded from below as follows
\[
R_{\pi}^S(H, T) \geq C_1(K - 1) \frac{1}{\Delta} \log \left( \frac{C_2 K^2}{T} \right),
\]
where \(C_1\) and \(C_2\) are the positive constants that were introduced in Theorem 1.

In the proof of Proposition 1 we show that, in fact, no admissible policy can guarantee any performance gain based on additional information even when the mappings \(\phi_k\) are known to be linear with an unknown parameter. We next describe the main ideas of the proof for the case of two arms. Denote by \(S' = S'(\Delta, \sigma^2, \hat{\sigma}^2)\), the class that includes pairs of allowed reward distribution profiles \((\nu, \nu')\), as described in §2 with the additional condition that \(\phi_k(\cdot) = \alpha_k \times (\cdot)\) for \(\alpha_k \in \{\alpha_1, \alpha_2\}\) for some \(\alpha_2 > \alpha_1\).

Note that \(S' \subset S\), which implies \(R_{\pi}^S(H, T) \geq R_{\pi}^{S'}(H, T)\). We assume that the decision maker is a priori informed that rewards from the first arm follow a normal distribution with a mean that is either \(\hat{\mu} - \Delta\) (according to \(\nu^{(1)}\)) or \(\hat{\mu} + \Delta\) (according to \(\nu^{(2)}\)), and the second arm is known to generate rewards with normal distribution with mean \(\hat{\mu}\). We set \(\hat{\mu}\) such that \(\frac{\hat{\mu} + \Delta}{\alpha_2} = \frac{\hat{\mu} - \Delta}{\alpha_1}\) with probability one in both cases. For the case \((\nu^{(1)}, \nu'^{(1)})\), an unbiased estimator for the mean reward of the second arm is constructed based on auxiliary observations \(\nu'^{(1)}\) and \(\nu'^{(2)}\) such that \(Y_{2,t} = \frac{\hat{\mu} + \Delta}{\alpha_2} = \frac{\hat{\mu} - \Delta}{\alpha_1}\) with probability one in both cases. In the case \((\nu^{(2)}, \nu'^{(2)})\), this unbiased estimator takes the form \(\alpha_2 \cdot Y_{2,t}\). Since auxiliary observations have the same distribution in both cases these observations are uninformative for the purpose of identifying the true reward distribution.

7.2 Leveraging auxiliary information flows for estimating the gaps \(\Delta_k\)

One of the challenges in the MAB literature is that one does not always have access to the true value of the gap parameter \(\Delta\). Hence, one might wonder whether auxiliary information flows could be leveraged to mitigate the effects of misspecification compared to the traditional setting without any auxiliary observations. For example, one might use the auxiliary observations to estimate the gaps \(\Delta_k\) along the path and use these estimates as inputs to their MAB policy in real time. To demonstrate how this can be done, we consider the \(\epsilon\)-greedy policy that directly uses the gap parameter, and evaluate its performance when it uses estimates maintained from auxiliary observations. We define:
\[
\bar{Z}_{k,t} := \begin{cases} 
\left( \sum_{s=1}^{t} Z_{k,s} \left( \sum_{s=1}^{t} h_{k,s} \right)^{-1} \right)^{-1} & \text{if } \sum_{s=1}^{t} h_{k,s} > 0 \\
0 & \text{o.w.} 
\end{cases}, \quad \hat{\Delta}_{k,t} := \max_{k' \in \mathcal{K}} \bar{Z}_{k',t} - \bar{Z}_{k,t}.
\]
In the following variation of the $\epsilon_t$-greedy policy we replace the parameter $\Delta$ with estimates $\{\hat{\Delta}_k\}$ that are derived based on auxiliary observations.

$\epsilon_t$-greedy with adaptive exploration and estimated gaps. 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 $h_t$ and $Z_t$, and update virtual time indexes for all $k \in K$:
      \[\tau_{k,t} = (\tau_{k,t-1} + 1) \cdot \exp \left( \frac{h_{k,t} \hat{\Delta}_{k,t}^2}{c \hat{\sigma}^2} \right)\]
   
   b) With probability $\min \{1, \sum_{k' = 1}^{K} \frac{c \sigma^2}{\hat{\Delta}_{k',t}} \}$ select an arm at random: (exploration)
      \[\pi_t = k \quad \text{with probability} \quad \frac{\tau_{-1}^{-1} \hat{\Delta}_{k,t}^{-2}}{\sum_{k' = 1}^{K} \tau_{k',t}^{-1} \hat{\Delta}_{k',t}^{-2}}, \quad \text{for all} \quad k \in K\]
      
      Otherwise, select an arm with the highest estimated reward: (exploitation)
      \[\pi_t = \arg \max_{k \in K} \bar{X}_{k,n_{k,t}}\]
   
   c) Receive and observe a reward $X_{\pi_t,t}$.

Proposition 2 (Near optimality under stationary information flows) Let $\pi$ be $\epsilon_t$-greedy with adaptive exploration and estimated gaps tuned by $c > 0$. If $h_{k,t}$'s are i.i.d. Bernoulli random variables with parameter $\lambda > 0$ then, for every $T \geq 1$ and $K \geq 2$:

\[
\mathbb{E}_H [R_S^\pi(H, T)] \leq \max_{k \in K} \Delta_k \cdot \left( \frac{2^7 \hat{\sigma}^2}{\lambda \Delta^2} \log \left( \frac{2^7 \hat{\sigma}^2}{\lambda \Delta^2} \right) + 2K + 10K/\lambda \right) + \sum_{k \in K \setminus \{k^*\}} \left( \frac{32c^2 \sigma^2 \hat{\sigma}^2}{\lambda \Delta^3_k} + \frac{64 \hat{\sigma}^2}{\Delta_k \lambda} \right).
\]

Proposition 2 states that when information arrival process is stationary, $\epsilon_t$-greedy with adaptive exploration and estimated gaps guarantees regret of $O(\lambda^{-1} \log \lambda^{-1})$ independently of $T$, which implies constant regret when $\lambda$ is large enough (the optimal regret rate in this setting scales with $\log \lambda^{-1}$; see discussion in §3.1.1 and Corollary 1 in Appendix C.1).

Proposition 3 (Near optimality under diminishing information flows) Let $\pi$ be $\epsilon_t$-greedy with adaptive exploration and estimated gaps, tuned by $c > 0$. If $h_{k,t}$'s are random variables such that for some $\kappa > \left( \frac{20 \Delta^2}{\bar{\sigma}^2} \lor 128 \lor 16c \right)$ one has $\mathbb{E} \left[ \sum_{s=1}^{t} h_{k,s} \right] = \left\lfloor \frac{\hat{\sigma}^2 k}{22 \lambda^2} \log t \right\rfloor$ for each arm $k \in K$ and each time step $t$. Then, for every $T \geq 1$ and $K \geq 2$:

32
\[ \mathbb{E}_H [R_S^T(H, T)] \leq \max_{k \in K} \Delta_k \left( K \frac{20\Delta^2}{(\sigma^2 \kappa - 20\Delta^2)} + 2K \frac{128}{(\kappa - 128)} \right) + \sum_{k \in K \setminus \{k^*\}} \left( \frac{16c^2 \sigma^2}{\Delta_k (\kappa - 16c)} + \frac{128\Delta^2 \Delta_k}{(\kappa \Delta^2_k - 32\Delta^2)} \right). \]

Recalling the discussion in §3.1.1, Proposition 3 establishes that when the information arrivals diminish with \( \kappa \) that is large enough, \( \epsilon_t \)-greedy with adaptive exploration and estimated gaps guarantees constant regret and therefore rate-optimal. We further note that in cases of interest \( \epsilon_t \)-greedy with adaptive exploration and estimated gaps may outperform a naive adaptation of \( \epsilon_t \)-greedy that uses an accurate gap parameter (but not virtual time indexes); for more details see numerical analysis in Appendix D.5.

### 7.3 Reactive information flows

We next discuss some implications of endogenous information flows that may depend on past decisions. The detailed formulation and analysis that support this discussion appear in Appendix C.6. For the sake of concreteness, we assume that information flows are polynomially proportional (decreasing or increasing) to the number of times various arms were selected. For some global parameters \( \gamma > 0 \) and \( \omega \geq 0 \) we assume that for each \( t \) and \( k \), the number of auxiliary observations collected up to time \( t \) on arm \( k \) is as follows:

\[ \sum_{s=1}^{t} h_{k,s} = \left[ \rho_k \cdot \tilde{n}_{k,t} + \sum_{j \in K \setminus \{k\}} \alpha_{k,j} \tilde{n}_{j,t} \right], \tag{4} \]

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 information arrivals of arm \( k \) on the past selections of arm \( k \), and for \( j \neq k \) the parameter \( \alpha_{k,j} \geq 0 \) captures the dependence of information arrivals of arm \( k \) on the past selections of arm \( j \). While the structure of (4) introduces some limitation on the impact past decisions may have on information flows, it still captures many types of dependencies. For example, when \( \gamma = 0 \), information arrivals are decoupled across arms (selecting an action can impact only future information on that action), but when \( \gamma > 0 \), a selection of a certain action may impact future information arrivals on all actions.

A key driver in the regret complexity is the relation between the number of times an arm is pulled, \( \tilde{n}_{k,t-1} \), and the total number observations of that arm, \( n_{k,t} = \tilde{n}_{k,t-1} + \sum_{s=1}^{t} \tilde{\sigma}^2 h_{k,t} \). 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 stochastic MAB literature; as this happens when \( \omega < 1 \), we next focus on the case \( \omega \geq 1 \). However, when pulling an arm increases information arrival rates to other arms, constant regret is achievable by a myopic policy (see Proposition 4 in Appendix C.6). In addition, when information flows are decoupled across arms, we further establish an improved upper bound under the class defined by (4) with \( \gamma = 0 \) and \( \omega \geq 1 \), showing that regret of order \((\log T)^{1/\omega}\) is achievable (see Proposition 5 in Appendix C.6).
7.4 Additional avenues for future research

While in §7.3 we discuss the optimality of our approach under a class of information arrival processes that can depend on the history of decisions and observations, the characterization of minimax regret rates and optimal policy design for general endogenous information flows is a challenging open problem.

Despite the impossibility (in general) of performance gain when the mappings $\phi_k$ is unknown, it is an open problem to potentially characterize classes of mappings under which one could adapt to unknown mapping $\phi_k$ while guaranteeing performance improvement based on auxiliary information.

We finally note that our work relates to settings with partial monitoring (Cesa-Bianchi et al. 2006), where agents do not observe payoff, but only some limited information about it. An interesting analogue would be to study adaptive policies with access to any combination of such feedback forms.
A  Proofs of Main results

A.1 Preliminaries

In the proofs, we use the following notation unless it is stated otherwise. Let \( \tilde{n}_{k,t} \) be the number of times arm \( k \) is pulled by the policy up to time \( t \):
\[
\tilde{n}_{k,t} := \sum_{s=1}^{t} \mathbb{1}\{\pi_s = k\}.
\]

Let \( n_{k,t} \) be the weighted number of samples observed from arm \( k \) up to time \( t \):
\[
n_{k,t} := \sum_{s=1}^{t-1} \mathbb{1}\{\pi_s = k\} + t \sum_{s=1}^{t} \frac{1}{\sigma^2} h_{k,s}.
\]

Denote by \( \bar{X}_{k,n_{k,t}} \) the weighted empirical average of the \( k \)th arm reward after \( n_{k,t} \) observations:
\[
\bar{X}_{k,n_{k,t}} := \frac{\sum_{s=1}^{t} \frac{1}{\sigma^2} \mathbb{1}\{\pi_s = k\} X_{k,s} + t \sum_{s=1}^{t} \frac{1}{\sigma^2} Z_{k,s}}{\sum_{s=1}^{t-1} \mathbb{1}\{\pi_s = k\} + t \sum_{s=1}^{t} \frac{1}{\sigma^2} h_{k,s}}.
\]

For any policy \( \pi \) and profiles \( \nu \) and \( \nu' \), let \( \mathbb{P}_{(\nu,\nu')}, \mathbb{E}_{(\nu,\nu')}, \text{ and } \mathbb{R}_{(\nu,\nu')} \) denote the probability, expectation, and regret when rewards are distributed according to \( \nu \), and auxiliary observations are distributed according to \( \nu' \).

A.2 Proof of Theorem 1

Step 1 (Notations and definitions). For \( m, q \in \{1, \ldots, K\} \) define the distribution profiles \( \nu^{(m,q)} \):
\[
\nu^{(m,q)}_k = \begin{cases} 
\mathcal{N}(0, \sigma^2) & \text{if } k = m \\
\mathcal{N}(\pm \Delta, \sigma^2) & \text{if } k \neq m \\
\mathcal{N}(0, \sigma^2) & \text{o.w.}
\end{cases}
\]

For example, for \( m = 1 \), one has
\[
\nu^{(1,1)} = \begin{pmatrix} 
\mathcal{N}(0, \sigma^2) \\
\mathcal{N}(\Delta, \sigma^2) \\
\mathcal{N}(\Delta, \sigma^2) \\
\mathcal{N}(\Delta, \sigma^2)
\end{pmatrix}, \quad \nu^{(1,2)} = \begin{pmatrix} 
\mathcal{N}(0, \sigma^2) \\
\mathcal{N}(\Delta, \sigma^2) \\
\mathcal{N}(\Delta, \sigma^2) \\
\mathcal{N}(\Delta, \sigma^2)
\end{pmatrix}, \quad \nu^{(1,3)} = \begin{pmatrix} 
\mathcal{N}(0, \sigma^2) \\
\mathcal{N}(\Delta, \sigma^2) \\
\mathcal{N}(\Delta, \sigma^2) \\
\mathcal{N}(\Delta, \sigma^2)
\end{pmatrix}, \quad \ldots, \quad \nu^{(1,K)} = \begin{pmatrix} 
\mathcal{N}(0, \sigma^2) \\
\mathcal{N}(\Delta, \sigma^2) \\
\mathcal{N}(\Delta, \sigma^2) \\
\mathcal{N}(\Delta, \sigma^2)
\end{pmatrix}.
\]

Similarly, assume that the auxiliary information \( Y_{k,t} \) is distributed according to \( \hat{\nu}_k^{(m,q)} \) and hence we use the notation \( \mathbb{P}_\nu, \mathbb{E}_\nu, \text{ and } \mathbb{R}_\nu \) instead of \( \mathbb{P}_{(\nu,\nu')}, \mathbb{E}_{(\nu,\nu')}, \text{ and } \mathbb{R}_{(\nu,\nu')} \).
Step 2 (Lower bound decomposition). We note that

\[
R_\Delta^0(H, T) \geq \max_{m,q \in \{1, \ldots, K\}} \left\{ \frac{1}{K} \sum_{m=1}^{K} \left( R_\Delta^\pi(H, T) \right) \right\} \geq \frac{1}{K} \sum_{m=1}^{K} \max_{q \in \{1, \ldots, K\}} \left\{ R_\Delta^\pi(H, T) \right\}.
\] (5)

Step 3 (A naive lower bound for \( \max_{q \in \{1, \ldots, K\}} \{ R_\Delta^\pi(H, T) \} \)). We note that

\[
\max_{q \in \{1, \ldots, K\}} \left\{ R_\Delta^\pi(m,q)(H, T) \right\} \geq R_\Delta^\pi(m,m)(H, T) = \Delta \cdot \sum_{k \in K \setminus \{m\}} E_{\nu(m,m)}[\hat{n}_{k,T}],
\] (6)

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 4, for any \( q \neq m \), one has

\[
KL(\nu_t^{(m,m)}, \nu_t^{(m,q)}) = \frac{2\Delta^2}{\sigma^2} \cdot E_{\nu(m,m)}[n_{q,t}] = \frac{2\Delta^2}{\sigma^2} \left( E_{\nu(m,m)}[\hat{n}_{q,t-1}] + \sum_{s=1}^{t} \frac{\sigma^2}{\sigma^2} h_{q,s} \right).
\] (7)

One obtains:

\[
\max_{q \in \{1, \ldots, K\}} \left\{ R_\Delta^\pi(m,q)(H, T) \right\} \geq \frac{1}{K} R_\Delta^\pi(m,m)(H, T) + \frac{1}{K} \sum_{q \in K \setminus \{m\}} R_\Delta^\pi(m,q)(H, T)
\]
\[
\geq \frac{\Delta}{K} \sum_{t=1}^{T} \sum_{k \in K \setminus \{m\}} P_{\nu(m,m)}(\pi_t = k) + \frac{\Delta}{K} \sum_{q \in K \setminus \{m\}} \sum_{t=1}^{T} P_{\nu(m,q)}(\pi_t \neq q)
\]
\[
= \frac{\Delta}{K} \sum_{t=1}^{T} \sum_{q \in K \setminus \{m\}} \left( P_{\nu(m,m)}(\pi_t = q) + P_{\nu(m,q)}(\pi_t \neq q) \right)
\]
\[
\geq \frac{\Delta}{2K} \sum_{t=1}^{T} \sum_{q \in K \setminus \{m\}} \exp(-KL(\nu_t^{(m,m)}, \nu_t^{(m,q)}))
\]
\[
\geq \frac{\Delta}{2K} \sum_{t=1}^{T} \sum_{q \in K \setminus \{m\}} \exp \left[ -\frac{2\Delta^2}{\sigma^2} \left( E_{\nu(m,m)}[\hat{n}_{q,t-1}] + \sum_{s=1}^{t} \frac{\sigma^2}{\sigma^2} h_{q,s} \right) \right]
\]
\[
\geq \frac{\Delta}{2K} \sum_{t=1}^{T} \sum_{q \in K \setminus \{m\}} \exp \left[ -\frac{2\Delta^2}{\sigma^2} \left( E_{\nu(m,m)}[\hat{n}_{q,t}] + \sum_{s=1}^{t} \frac{\sigma^2}{\sigma^2} h_{q,s} \right) \right]
\]
\[
= \sum_{q \in K \setminus \{m\}} \frac{\Delta}{2K} \exp \left( -\frac{2\Delta^2}{\sigma^2} E_{\nu(m,m)}[\hat{n}_{q,T}] \right) \sum_{t=1}^{T} \exp \left( -\frac{2\Delta^2}{\sigma^2} \sum_{s=1}^{t} h_{q,s} \right),
\] (8)

where (a) follows from Lemma 3, (b) holds by (7), and (c) follows from \( \hat{n}_{q,T} \geq \hat{n}_{q,t-1} \) for \( t \in T \).

Step 5 (Unifying the lower bounds in steps 3 and 4). Using (6), and (8), we establish
\[
\max_{q \in \{1, \ldots, K\}} \left\{ R_{\nu(m, q)}^T(\mathbf{H}, T) \right\} \geq \frac{\Delta}{2} \sum_{k \in K \setminus \{m\}} \left( \mathbb{E}_{\nu(m, m)}[\hat{\nu}_k, T] + \exp \left( \frac{-2\Delta^2 \cdot \mathbb{E}_{\nu(m, m)}[\hat{\nu}_k, T]}{2K} \right) \sum_{t=1}^T \exp \left( -\frac{2\Delta^2}{\sigma^2} \sum_{s=1}^t h_{k,s} \right) \right)
\]

\[
\geq \frac{\Delta}{2} \sum_{k \in K \setminus \{m\}} \min_{\nu \geq 0} \left( x + \frac{\exp \left( \frac{-2\Delta^2}{\sigma^2} \cdot x \right) \sum_{t=1}^T \exp \left( -\frac{2\Delta^2}{\sigma^2} \sum_{s=1}^t h_{k,s} \right)}{2} \right)
\]

\[
\geq \frac{\sigma^2}{4\Delta} \sum_{k \in K \setminus \{m\}} \log \left( \frac{\Delta^2 \sum_{t=1}^T \exp \left( -\frac{2\Delta^2}{\sigma^2} \sum_{s=1}^t h_{k,s} \right)}{\sigma^2 K} \right),
\]

where (a) follows from \( x + \gamma e^{-\kappa x} \geq \frac{\log \gamma \kappa}{\kappa} \) for \( \gamma, \kappa, x > 0 \). (Note that the function \( x + \gamma e^{-\kappa x} \) 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 \([5]\), and \([9]\). \( \square \)

### A.3 Proof of Theorem 2

We adapt the proof of the upper bound for Thompson sampling with Gaussian priors in \cite{AgrawalGoyal2013a}.

**Step1 (Notations and definitions).** For every suboptimal arm \( k \), we consider three parameters \( x_k, y_k, \) and \( u_k \) such that \( \mu_k < x_k < y_k < u_k < \mu_{k^*} \). We will specify these three parameters at the end of the proof. Also, define the events \( E_{k,t}^\mu \) and \( E_{k,t}^\theta \) to be the events on which \( \bar{X}_{k,\pi_k, t} \leq x_k \) and \( \theta_{k,t} \leq y_k \), respectively. In words, the events \( E_{k,t}^\mu \) and \( E_{k,t}^\theta \) happen when the estimated mean reward, and the sample mean reward do not deviate from the true mean, respectively. Define the history \( \mathcal{H}_t := \{ \{X_{\pi_s, s}\}_{s=1}^{t-1}, \{\pi_s\}_{s=1}^{t-1}, \{Z_s\}_{s=1}^{t-1}, \{h_s\}_{s=1}^{t-1} \} \) for all \( t = 1, \ldots, T \). Finally define

\[
p_{k,t} = P_{(\nu, \nu')}(\theta_{k^*, t} > y_k \mid \mathcal{H}_t).
\]

**Step2 (Preliminaries).** We will make use of the following lemmas from \cite{AgrawalGoyal2013a} throughout this proof. The proof of the lemmas are skipped since they are simple adaptation to our setting with auxiliary information arrival.

**Lemma 1** For any suboptimal arm \( k \),

\[
\sum_{t=1}^T P_{(\nu, \nu')}\left\{ \pi_t = k, E_{k,t}^\mu, E_{k,t}^\theta \right\} \leq \sum_{j=0}^{T-1} \mathbb{E} \left[ \frac{1 - p_{k,t_j+1}}{p_{k,t_j+1}} \right],
\]

where \( t_0 = 0 \) and for \( j > 0 \), \( t_j \) is the time step at which the optimal arm \( k^* \) is pulled for the \( j \)th time.
Proof. The proof can be found in the analysis of Theorem 1 in [Agrawal and Goyal (2013a)]. However, we bring the proof for completeness.

\[
\sum_{t=1}^{T} \Pr_{\pi}^{(\nu,\nu')} \{ \pi_t = k, E_{k,t}^{\mu}, E_{k,t}^{\theta} \} = \sum_{t=1}^{T} \mathbb{E}_{\pi}^{(\nu,\nu')} \left[ \Pr \left\{ \pi_t = k, E_{k,t}^{\mu}, E_{k,t}^{\theta} \mid H_t \right\} \right]
\]

\[
\leq \sum_{t=1}^{T} \mathbb{E}_{\pi}^{(\nu,\nu')} \left[ \frac{1 - P_{k,t}}{P_{k,t}} \Pr \left\{ \pi_t = 1, E_{k,t}^{\mu}, E_{k,t}^{\theta} \mid H_t \right\} \right]
\]

\[
= \sum_{t=1}^{T} \mathbb{E}_{\pi}^{(\nu,\nu')} \left[ \mathbb{E} \left[ \frac{1 - P_{k,t}}{P_{k,t}} \mathbbm{1} \left\{ \pi_t = 1, E_{k,t}^{\mu}, E_{k,t}^{\theta} \right\} \mid H_t \right] \right]
\]

\[
= \sum_{t=1}^{T} \mathbb{E}_{\pi}^{(\nu,\nu')} \left[ \frac{1 - P_{k,t}}{P_{k,t}} \mathbbm{1} \left\{ \pi_t = 1, E_{k,t}^{\mu}, E_{k,t}^{\theta} \right\} \right]
\]

\[
\leq \sum_{j=1}^{T-1} \mathbb{E}_{\pi}^{(\nu,\nu')} \left[ \frac{1 - P_{k,t_j+1}}{P_{k,t_j+1}} \sum_{t=t_j+1}^{t_j+1} \mathbbm{1} \left\{ \pi_t = 1 \right\} \right]
\]

\[
= \sum_{j=0}^{T-1} \mathbb{E} \left[ \frac{1 - P_{k,t_j+1}}{P_{k,t_j+1}} \right],
\]

where (a) follows from Lemma 1 in [Agrawal and Goyal (2013a)].

Lemma 2 (Agrawal and Goyal 2013a, Lemma 2) For any suboptimal arm \( k \),

\[
\sum_{t=1}^{T} \mathbb{E}_{\pi}^{(\nu,\nu')} \{ \pi_t = k, E_{k,t}^{\mu} \} \leq 1 + \frac{\sigma^2}{(x_k - \mu_k)^2}.
\]

Step 3 (Regret decomposition). Fix a profile \( \nu \). For each suboptimal arm \( k \), we will decompose the number of times it is pulled by the policy as follows and upper bound each term separately:

\[
\mathbb{E}_{(\nu,\nu')}^{\pi} \{ \tilde{n}_{k,T} \} = \sum_{t=1}^{T} \mathbb{E}_{(\nu,\nu')}^{\pi} \left\{ \pi_t = k, E_{k,t}^{\mu}, E_{k,t}^{\theta} \right\}^{J_{k,1}} + \sum_{t=1}^{T} \mathbb{E}_{(\nu,\nu')}^{\pi} \left\{ \pi_t = k, E_{k,t}^{\mu}, E_{k,t}^{\theta} \right\}^{J_{k,2}} + \sum_{t=1}^{T} \mathbb{E}_{(\nu,\nu')}^{\pi} \left\{ \pi_t = k, E_{k,t}^{\mu} \right\}^{J_{k,3}}.
\]

(10)

Step 4 (Analysis of \( J_{k,1} \)). Given Lemma 1, \( J_{k,1} \) can be upper bounded by analyzing \( \frac{1 - p_{k,t_j+1}}{p_{k,t_j+1}} \):

\[
p_{k,t_j+1} = \mathbb{P} \left\{ N \left( \bar{X}_{k,n_k}, c \frac{\sigma^2}{j+1} \right) > y_k \right\}
\]

\[
\geq \mathbb{P} \left\{ N \left( \bar{X}_{k,n_k}, c \frac{\sigma^2}{j+1} \right) > y_k \mid \bar{X}_{k,n_k} \geq u_k \right\} \cdot \mathbb{P} \left\{ \bar{X}_{k,n_k} \geq u_k \right\} =: Q_1 \cdot Q_2.
\]
First, we analyze $Q_1$:

$$Q_1 \geq \mathbb{P}\left\{ N\left(u_k, \frac{c\sigma^2}{j+1}\right) > y_k \right\} \geq 1 - \exp\left(-\frac{(j+1)(u_k - y_k)^2}{2\sigma^2}\right),$$

where the last inequality follows from Chernoff-Hoeffding bound in Lemma 5. The term $Q_2$ can also be bounded from below using Chernoff-Hoeffding in Lemma 5 bound as follows:

$$Q_2 \geq 1 - \exp\left(-\frac{j(\mu_1 - u_k)^2}{2\sigma^2}\right).$$

Define $\delta_k := \min\{(\mu_1 - u_k), (u_k - y_k)\}$. If $j \geq \frac{(c\vee 1)\sigma^2 \ln 4}{\delta_k^2}$ then, $p_{k,t+1} \geq 1/4$ and $1 - p_{k,t+1} \leq 2\exp(-j\delta_k^2/2\sigma^2(c \vee 1))$. Also, for $1 \leq j < \frac{(c\vee 1)\sigma^2 \ln 4}{\delta_k^2}$ we have $p_{k,t+1} \geq (1 - \exp(-\delta_k^2/2\sigma^2(c \vee 1)))^2$ and $1 - p_{k,t+1} \leq 1$. Finally, note that $p_{k,t+1} = \mathbb{P}\{N(0,c\sigma^2) > y_k\}$. Plugging these inequalities into Lemma 1, we obtain

$$J_{k,1} \leq \frac{16(c \vee 1)\sigma^2}{\delta_k^2} + \frac{(c \vee 1)\sigma^2 \ln 4}{\delta_k^2(1 - \exp(-\delta_k^2/2\sigma^2(c \vee 1)))^2} + \frac{1}{\mathbb{P}\{N(0,c\sigma^2) > y_k\}} - 1. \quad (11)$$

**Step 5 (Analysis of $J_{k,2}$).** First, we upper bound the following conditional probability

$$\mathbb{P}_{(\nu,\nu')}\left\{ \pi_t = k, E_{k,t}^\mu, E_{k,t}^\theta \mid \epsilon_{k,t} \right\} \leq \mathbb{P}_{(\nu,\nu')}\left\{ E_{k,t}^\mu, E_{k,t}^\theta \mid \epsilon_{k,t} \right\} \leq \mathbb{P}_{(\nu,\nu')}\left\{ E_{k,t}^\theta \mid E_{k,t}^\mu, \epsilon_{k,t} \right\} \leq \mathbb{P}\left\{ N\left(\bar{X}_{k,n_{k,t}}^{\mu} (n_{k,t} + 1)^{-1}, n_{k,t} \right) > y_k \mid \epsilon_{k,t} \right\} \leq \mathbb{P}\left\{ N\left(x_k^{\mu} (n_{k,t} + 1)^{-1}, n_{k,t} \right) > y_k \mid \epsilon_{k,t} \right\} \leq \exp\left(-\frac{n_{k,t}(y_k - x_k)^2}{2\sigma^2}\right), \quad (12)$$

where (a) follows from Chernoff-Hoeffding’s bound in Lemma 5. Define

$$n_{k,t}^\theta := \sum_{s=1}^{t} \frac{\sigma^2}{\sigma^2} h_{k,s} + \sum_{s=1}^{t-1} 1 \{ \pi_s = k, E_{k,s}^\mu, E_{k,s}^\theta \} \leq n_{k,t}.$$  

By (12), and the definition above, one obtains

$$\mathbb{P}_{(\nu,\nu')}\left\{ \pi_t = k, E_{k,t}^\mu, E_{k,t}^\theta \mid n_{k,t}^\theta \right\} \leq \exp\left(-\frac{n_{k,t}^\theta(y_k - x_k)^2}{2\sigma^2}\right)$$
for every $t \in T$. Note that by definition,
\[
n_{k,t}^\theta = n_{k,t-1}^\theta + \sigma^2 h_{k,t} + 1 \left\{ \pi_{t-1} = k, E_{k,t-1}^\mu, \tilde{E}_{k,t-1}^\theta \right\}.
\]

By the above two displays, we can conclude that
\[
\mathbb{E}_{(\nu,\nu')}^\pi \left[ n_{k,t}^\theta \mid n_{k,t-1}^\theta \right] = n_{k,t-1}^\theta + \frac{\sigma^2}{\alpha^2} h_{k,t} + \mathbb{P}_{(\nu,\nu')}^\pi \left\{ \pi_{t-1} = k, E_{k,t-1}^\mu, \tilde{E}_{k,t-1}^\theta \right\} \leq n_{k,t-1}^\theta + \frac{\sigma^2}{\alpha^2} h_{k,t} + e^{-\frac{\nu_{k,t-1}^\theta (y_k - x_k)^2}{2c\alpha^2}}.
\]

Now, we can apply Lemma 11 to the sequence $\left\{ n_{k,t}^\theta \right\}_{0 \leq t \leq T}$, and obtain
\[
\mathbb{E}_{(\nu,\nu')}^\pi \left[ n_{k,T+1}^\theta \mid n_{k,t}^\theta \right] \leq \log \exp \left( \frac{(y_k - x_k)^2}{2c\alpha^2} \right) \left( 1 + \sum_{t=1}^T \exp \left( -\frac{(y_k - x_k)^2}{2c\alpha^2} \sum_{s=1}^t h_{k,s} \right) \right) + \sum_{t=1}^T \frac{\sigma^2}{\alpha^2} h_{k,t}
\]
\[
\leq \frac{2c\alpha^2}{(y_k - x_k)^2} \log \left( 1 + \sum_{t=1}^T \exp \left( -\frac{(y_k - x_k)^2}{2c\alpha^2} \sum_{s=1}^t h_{k,s} \right) \right) + \sum_{t=1}^T \frac{\sigma^2}{\alpha^2} h_{k,t}.
\]

By this inequality and the definition of $n_{k,t}^\theta$, we have
\[
J_{k,2} = \sum_{t=1}^T \mathbb{P}_{(\nu,\nu')}^\pi \left\{ \pi_t = k, E_{k,t}^\mu, \tilde{E}_{k,t}^\theta \right\} \leq \frac{2c\alpha^2}{(y_k - x_k)^2} \log \left( 1 + \sum_{t=1}^T \exp \left( -\frac{(y_k - x_k)^2}{2c\alpha^2} \sum_{s=1}^t h_{k,s} \right) \right).
\]

**Step 6 (Analysis of $J_{k,3}$).** The term $J_{k,3}$ can be upper bounded using Lemma 2.

**Step 7 (Determining the constants).** Finally, let $x_k = \mu_k + \frac{\Delta_k}{T}$, $y_k = \mu_k + \frac{2\Delta_k}{T}$, and $u_k = \mu_k + \frac{3\Delta_k}{T}$.

Then, by putting (10), (11), (13), and Lemma 2 back together, the result is established.

---

### A.4 Proof of Theorem 3

Fix a profile $(\nu,\nu') \in \mathcal{S}$ and consider a suboptimal arm $k \neq k^*$. If arm $k$ is pulled at time $t$, then
\[
X_{k,n_k,t} + \sqrt{\frac{c\alpha^2 \log t}{n_k,t}} \geq X_{k^*,n_{k^*},t} + \sqrt{\frac{c\alpha^2 \log t}{n_{k^*},t}}.
\]

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

- $\mathcal{E}_{1,t} := \left\{ X_{k,n_k,t} \geq \mu_k + \sqrt{\frac{c\alpha^2 \log t}{n_k,t}} \right\}$,
- $\mathcal{E}_{2,t} := \left\{ X_{k^*,n_{k^*},t} \leq \mu_{k^*} - \sqrt{\frac{c\alpha^2 \log t}{n_{k^*},t}} \right\}$,
- $\mathcal{E}_{3,t} := \left\{ \Delta_k \leq 2 \sqrt{\frac{c\alpha^2 \log t}{n_k,t}} \right\}$.

To see why this is true, assume that all the above events fail. Then, we have...
\[ X_{k,n_k,t} + \sqrt{\frac{c\sigma^2 \log t}{n_{k,t}}} \leq \mu_k + 2 \sqrt{\frac{c\sigma^2 \log t}{n_{k,t}}} < \mu_k + \Delta_k = \mu_{k^*} < X_{k^*,n_{k^*,t}} + \sqrt{\frac{c\sigma^2 \log t}{n_{k^*,t}}}, \]

which is in contradiction with the assumption that arm \( k \) is pulled at time \( t \). 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 \( t \in \mathcal{T} \), one has

\[
E_{\pi}(\nu,\nu')[\tilde{n}_{k,T}] = E_{\pi}(\nu,\nu') \left[ \sum_{t=1}^{T} 1 \{ \pi_t = k \} \right] = E_{\pi}(\nu,\nu') \left[ \sum_{t=1}^{T} 1 \{ \pi_t = k, n_{k,t} \leq \hat{l}_{k,t} \} + 1 \{ \pi_t = k, n_{k,t} > \hat{l}_{k,t} \} \right] \leq E_{\pi}(\nu,\nu') \left[ \sum_{t=1}^{T} 1 \{ \pi_t = k, n_{k,t} \leq \hat{l}_{k,t} \} + 1 \{ \pi_t = k, n_{k,t} > \hat{l}_{k,t} \} \right] \leq E_{\pi}(\nu,\nu') \left[ \sum_{t=1}^{T} \left( \pi_t = k, \tilde{n}_{k,t-1} \leq \hat{l}_{k,t} - \sum_{s=1}^{t} \frac{\sigma^2}{\sigma^2} h_{k,s} \right) \right] + \sum_{t=1}^{T} \mathbb{P} \{ \pi_t = k, n_{k,t} > \hat{l}_{k,t} \} \leq \max_{1 \leq t \leq T} \left\{ \hat{l}_{k,t} - \sum_{s=1}^{t} \frac{\sigma^2}{\sigma^2} h_{k,s} \right\} + \sum_{t=1}^{T} \mathbb{P} \{ \pi_t = k, n_{k,t} > \hat{l}_{k,t} \}.
\]

Set the values of \( l_{k,t} \) and \( \hat{l}_{k,t} \) as follows:

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

where

\[
\tau_{k,t} := \sum_{s=1}^{t} \exp \left( \frac{\Delta_k}{4c\sigma^2} \sum_{r=s}^{t} h_{k,r} \right).
\]

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

\[
E_{\pi}(\nu,\nu')[\tilde{n}_{k,T}] \leq \max_{1 \leq t \leq T} \left\{ \hat{l}_{k,t} - \sum_{s=1}^{t} \frac{\sigma^2}{\sigma^2} h_{k,s} \right\} + \sum_{t=1}^{T} \mathbb{P} \{ \pi_t = k, E_{3,t}^c \} \leq \max_{1 \leq t \leq T} \left\{ \hat{l}_{k,t} - \sum_{s=1}^{t} \frac{\sigma^2}{\sigma^2} h_{k,s} \right\} + \sum_{t=1}^{T} \mathbb{P} \{ E_{1,t} \cup E_{2,t} \} \leq \max_{1 \leq t \leq T} \left\{ \hat{l}_{k,t} - \sum_{s=1}^{t} \frac{\sigma^2}{\sigma^2} h_{k,s} \right\} + \sum_{t=1}^{T} \frac{2}{t^{c/2}}, \tag{14}
\]

where the last inequality follows from Lemma 5 and the union bound. Plugging the value of \( \hat{l}_{k,t} \) into (14), one obtains
\[ \mathbb{E}_\pi^{\pi}(\nu, \nu') \left[ \hat{n}_{k,T} \right] \leq \max_{1 \leq t \leq T} \left\{ \frac{4c^2}{\Delta_k^2} \log \left( \sum_{s=1}^{t} \exp \left( \frac{\Delta_s^2}{4c^2} \sum_{\tau=s}^{t} h_{k,\tau} \right) - \frac{\Delta_k^2}{4c^2} \sum_{\tau=1}^{t} h_{k,\tau} \right) \right\} + \sum_{t=1}^{T} \frac{2}{t^{c/2}} \]

which concludes the proof. ■

### A.5 Proof of Theorem 4

Let \( t^* = \left\lceil \frac{Kc^2}{\Delta^2} \right\rceil \), and fix \( T \geq t^* \), \( K \geq 2 \), \( H \in \{0, 1\}^{K \times T} \), and profile \( (\nu, \nu') \in \mathcal{S} \).

**Step 1 (Decomposing the regret).** We decompose the regret of a suboptimal arm \( k \neq k^* \) as follows:

\[
\Delta_k \cdot \mathbb{E}_\pi^{\pi}(\nu, \nu') \left[ \sum_{t=1}^{T} \mathbb{1} \{ \pi_t = k \} \right] = \underbrace{\Delta_k \cdot \mathbb{E}_\pi^{\pi}(\nu, \nu') \left[ \sum_{t=1}^{T} \mathbb{1} \{ \pi_t = k, \text{ policy does exploration} \} \right]}_{J_{k,1}} + \underbrace{\Delta_k \cdot \mathbb{E}_\pi^{\pi}(\nu, \nu') \left[ \sum_{t=1}^{T} \mathbb{1} \{ \pi_t = k, \text{ policy does exploitation} \} \right]}_{J_{k,2}}. \tag{15}
\]

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:

\[
\tau_{k,t} = \sum_{s=1}^{t} \exp \left( \frac{\Delta_s^2}{c^2} \sum_{\tau=s}^{t} h_{k,\tau} \right) . \tag{16}
\]

**Step 3 (Analysis of \( J_{k,1} \)).** Let \( \bar{M} := \sum_{t=1}^{T} h_{k,t} \). For \( 0 \leq m \leq \bar{M} \), let \( t_m \) be the time step at which the \( m \)th auxiliary observation for arm \( k \) is received, that is,

\[
t_m = \begin{cases} 
\min \left\{ 1 \leq t \leq T \mid \sum_{s=1}^{t} h_{k,s} = m \right\} & \text{if } 0 \leq m \leq \bar{M} \\
T + 1 & \text{if } m = \bar{M} + 1
\end{cases}.
\]

Note that we dropped the index \( k \) in the definitions above for simplicity of notation. Also define \( \tau_{k,T+1} := (\tau_{k,T} + 1) \cdot \exp \left( \frac{h_{k,T} \Delta^2}{c^2} \right) \). One obtains
\[ J_{k,1} = \sum_{t=1}^{T} \min \left\{ \frac{c\sigma^2}{\Delta^2} \sum_{k'=1}^{K} \frac{1}{\tau_{k',t}}, 1 \right\} \cdot \frac{1}{\tau_{k,t}} \]
\[ \leq \sum_{t=1}^{T} \frac{c\sigma^2}{\Delta^2 \tau_{k,t}} \]
\[ = \sum_{m=0}^{M} \sum_{s=0}^{t_{m+1}-t_m-1} \frac{c\sigma^2}{\Delta^2 (\tau_{k,t_m} + s)} \]
\[ \leq \sum_{m=0}^{M} \frac{c\sigma^2}{\Delta^2} \log \left( \frac{\tau_{k,t_m} + t_{m+1} - t_m - 1 + 1/2}{\tau_{k,t_m} - 1/2} \right) \]
\[ \leq \sum_{m=0}^{M} \frac{c\sigma^2}{\Delta^2} \log \left( \frac{\exp \left( -\frac{\Delta^2}{c\sigma^2} \right) \cdot (\tau_{k,t_m+1} - 1/2)}{\tau_{k,t_m} - 1/2} \right) \]
\[ = \frac{c\sigma^2}{\Delta^2} \log \left( \frac{\tau_{k,\bar{m}+1} - 1/2}{\tau_{k,t_0} - 1/2} \right) - \sum_{t=1}^{T} \frac{\sigma^2}{\Delta^2} h_{k,t} \]
\[ \leq \frac{c\sigma^2}{\Delta^2} \log \left( \frac{\tau_{k,T+1} - 1/2}{1 - 1/2} \right) - \sum_{t=1}^{T} \frac{\sigma^2}{\Delta^2} h_{k,t} \]
\[ \leq \frac{c\sigma^2}{\Delta^2} \log \left( 2 \sum_{t=0}^{T} \exp \left( -\frac{\Delta^2}{c\sigma^2} \sum_{s=1}^{t} h_{k,s} \right) \right), \quad (17) \]

where (a), (b), and (c) follow from Lemma 9, the fact that \( \exp \left( -\frac{\Delta^2}{c\sigma^2} \right) \) \( \tau_{k,t_m} + t_{m+1,k} - t_m = \tau_{k,t_{m+1}} \), and (16), 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 \( t \geq t^* \). Fix \( t \geq t^* \). Then, \( n_{k,t} \) is a summation of independent Bernoulli r.v.’s. One has

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

The term \( \sum_{s=t^*}^{t-1} \frac{c\sigma^2}{\Delta^2 \tau_{k,s}} \) can be bounded from below similar to Step 3. Let \( \bar{M}_t = \sum_{s=1}^{t-1} h_{k,s} \) and \( M = \sum_{s=1}^{t^*} h_{k,s} \).

For \( M \leq m \leq \bar{M}_t \), 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,

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

One has that
\[
\sum_{s=t^*}^{t-1} \frac{c\sigma^2}{\Delta^2 t_{k,s}} = \sum_{m=M}^{M_t} \sum_{s=0}^{t_m,t-1} \frac{c\sigma^2}{\Delta^2 (t_{k,m,t} + s)} \\
\geq (a) \sum_{m=M}^{M_t} \frac{c\sigma^2}{\Delta^2} \log \left( \frac{t_{k,m,t} + t_{m+1,t} - t_{m,t}}{t_{k,m,t}} \right) \\
= \sum_{m=M}^{M_t} \frac{c\sigma^2}{\Delta^2} \log \left( \exp \left( \frac{\Delta^2}{c\sigma^2} t_{k,m+1,t} \right) \right) \\
= \frac{c\sigma^2}{\Delta^2} \log \left( \frac{t_{k,t-1}}{t_{k,t^*}} \right) - \sum_{s=t^*+1}^{t-1} \frac{\sigma^2}{\Delta^2} h_{k,s} \\
= \frac{c\sigma^2}{\Delta^2} \log \left( \frac{t_{k,t-1}}{t^*} \right) - \sum_{s=1}^{t-1} \frac{\sigma^2}{\Delta^2} h_{k,s},
\]  

(19)

where (a) follows from Lemma 9. Putting together (18) and (19), one obtains

\[
\mathbb{E} [n_{k,t}] \geq \frac{c\sigma^2}{\Delta^2} \log \left( \frac{t_{k,t-1}}{t^*} \right) := x_t.
\]

(20)

By Bernstein’s inequality in Lemma 6, we have

\[
\mathbb{P} \left\{ n_{k,t} \leq \frac{x_t}{2} \right\} \leq \exp \left( -\frac{x_t}{10} \right) = \exp \left( -\frac{t^*}{\tau_{k,t-1}} \right) \leq \exp \left( -\frac{t^*}{\Delta^2} \right) \leq \exp \left( -\frac{t^*}{\Delta^2} \right),
\]

(21)

where the last inequality follows from (16).

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

\[
\sum_{k \in K \setminus k^*} J_{k,2} \leq t^* - 1 + \sum_{k \in K \setminus k^*} \sum_{t=t^*}^{T} \mathbb{P} \left\{ X_{k,n_{k,t}} > X_{k^*,n_{k^*,t}} \right\}.
\]

(22)

We upper bound each summand as follows:

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

(23)

We next bound the first term on the right-hand side of the above inequality; the second term can be treated similarly. One obtains
\[
\mathbb{P}\left\{ \bar{X}_{k,n_k,t} > \mu_k + \frac{\Delta_k}{2} \right\} \leq \mathbb{P}\left\{ \bar{X}_{k,n_k,t} > \mu_k + \frac{\Delta_k}{2} \mid n_k,t > \frac{x_t}{2} \right\} + \mathbb{P}\left\{ n_k,t \leq \frac{x_t}{2} \right\}
\]

\[
\leq \exp\left( -\frac{\Delta_k^2 x_t}{16\sigma^2} \right) + \left( \frac{t^*}{t-1} \right)^{-\frac{c}{16} \sigma^2 \frac{t}{t-1}}
\]

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

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

where, (a) holds by the conditional probability definition, (b) follows from Lemma 5 together with (21), and (c) holds by (16). Putting together (15), (17), (22), (23), and (24), the result is established. □

A.6 Proof of Theorem 5

For the sake of brevity, we only prove the second part of the statement; the first part is a simple adaptation of the proof in Goldenshluger and Zeevi (2013). Similarly, the proof of the second statement adapts ideas from the proof of Theorem 1 in Goldenshluger and Zeevi (2013) to our setting as well. In what follows we will highlight the differences.

Similar to (20) in the proof of Theorem 4, we can show that

\[
\mathbb{E}[\bar{n}_{k,t}] \geq \frac{t^* - 1}{K} + \frac{2}{q} \log \left( \frac{\tau_{k,t-1}}{t^*} \right) \geq \frac{2}{q} \log \left( \frac{\tau_{k,t-1}}{t^*} \right),
\]

where the last inequality follows from \( t^* \geq \frac{2K}{q} \) and \( q \leq 1/5 \). Define the event \( \mathcal{E}_t := \{ \exists k : \bar{n}_{k,t} < \frac{1}{q} \log \left( \frac{\tau_{k,t-1}}{t^*} \right) \} \). By Bernstein’s inequality in Lemma 6 we have

\[
\mathbb{P}\{ \mathcal{E}_t \} \leq K \exp\left( -\frac{\log \left( \frac{\tau_{k,t-1}}{t^*} \right)}{5q} \right) \leq K \cdot \left( \frac{1}{\tau_{k,t-1}} \right)^{1/5q} \leq K \cdot \left( \frac{1}{t-1} \right)^{1/5q}.
\]

One can show that the regret incurred after \( t^* \) given that at least one of the events \( \mathcal{E}_t, t = t^*, \ldots, T \) is constant with respect to \( T \):

\[
\mathbb{E}\left[ \sum_{t=t^*}^{T} (\beta_1 - \beta_2)^\top W_t \mid \mathbb{1}\{\pi_t \neq \pi^*_t\} \cdot \mathbb{1}\{\mathcal{E}_t\} \right] \leq \mathbb{E}\left[ \sum_{t=t^*}^{T} (\beta_1 - \beta_2)^\top W_t \mid \mathbb{1}\{\mathcal{E}_t\} \right] \leq \sum_{t=t^*}^{T} Kb \sqrt{1 + dr^2} \left( \frac{1}{t-1} \right)^{1/5q} \leq C_{14},
\]

where \( C_{14} \) is independent of \( T \). Hence, we can analyze the regret conditional on the event that none of the events \( \mathcal{E}_t \) happen. Define \( \mathcal{F}_k := \{ t \in T \mid \pi_t = k, G_t \leq \frac{2}{q} \sum_{k'=1}^{K} \tau_{k',t}^{-1} \} \) to be the time steps that the
policy explores arm $k$, and let $\bar{T} = \bar{T}_1 \cup \bar{T}_2$. Then, one has:

$$
\mathcal{R}_\pi (\mathbf{H}, T) \leq E \left[ \sum_{t=1}^{n} (\beta_1 - \beta_2)^T W_t \cdot 1 \{ \pi_t \neq \pi_t^* \} \cdot 1 \{ \mathcal{E}_t \} \right] + E \left[ \sum_{t=1}^{n} (\beta_1 - \beta_2)^T W_t \cdot 1 \{ \pi_t \neq \pi_t^* \} \cdot 1 \{ \mathcal{E}_t^c \} \right]
$$

$$
\leq C_{14} + E \left[ \sum_{t=1, t\in \bar{T}}^{n} (\beta_1 - \beta_2)^T W_t \cdot 1 \{ \pi_t \neq \pi_t^* \} \cdot 1 \{ \mathcal{E}_t \} \right] + E \left[ \sum_{t=1, t\not\in \bar{T}}^{n} (\beta_1 - \beta_2)^T W_t \cdot 1 \{ \pi_t \neq \pi_t^* \} \cdot 1 \{ \mathcal{E}_t \} \right]
$$

$$
\leq C_{14} + K b \sqrt{\lambda_{\text{max}}(Q) E[|\bar{T}|]} + E \left[ \sum_{t=1}^{n} (\beta_1 - \beta_2)^T W_t \cdot 1 \{ \pi_t \neq \pi_t^* \} \cdot 1 \{ \mathcal{E}_t \} \right]. \quad (25)
$$

One may bound the second term in the last inequality by observing that $|\bar{T}| = |\bar{T}_1| + |\bar{T}_2| = \hat{n}_{1,T} + \hat{n}_{2,T}$, which can be upper bounded similar to (17) as follows:

$$
|\bar{T}| \leq t^* - 1 + \sum_{k=1}^{K} 2 \log \left( \frac{1}{t^* - 1/2} \sum_{t=t^*+1}^{T} \exp \left( \frac{1}{2} \sum_{s=1}^{t} h_{k,t} \right) \right) + 1.
$$

The third term in (25) can be upper bounded in a manner similar to [Goldenshluger and Zeevi, 2013], except for the term $J_2^{(1)}(t)$ in their analysis. When $\alpha > 1$, this term will be bounded as follows:

$$
J_2^{(1)}(t) \leq \left( c_1 K_{a} d^2 \lambda_{\pi}^{-2} \sigma^2 (1 + r^2 d) \max\{1, r^2\} t^{-1} \right)^{1+\alpha/2},
$$

where $c_1$ is an absolute constant and $K_a = \frac{(\alpha/2)^{\alpha/2}}{1-2^{-\alpha}} + \frac{\Gamma(\alpha/2)}{2 \ln 2}$ with $\Gamma(\cdot)$ as the Gamma-function. This modification allows one to bound the third term in (25) by a constant that depends on the parameters of the class $\mathcal{P}_w, \sigma^2, b, r, \alpha$ and $d$. 

\section{A.7 Proof of Proposition 1}

Denote by $\mathcal{S}' = \mathcal{S}'(\Delta, \sigma^2, \bar{\sigma}^2)$, the class that includes pairs of allowed reward distribution profiles $(\nu, \nu')$, as described in §2 with an extra condition that $\phi_k(\cdot) = \alpha_k \times (\cdot)$ for $\alpha_k \in \{\alpha_1, \alpha_2\}$ for some $\alpha_2 > \alpha_1$. Note that $\mathcal{S}' \subseteq \mathcal{S}$, which implies $\mathcal{R}_\pi^S(\mathbf{H}, T) \geq \mathcal{R}_\pi^{S_0}(\mathbf{H}, T)$. Let $\hat{\mu}$ be the solution to the equation $\alpha_1 (\mu + \Delta) = \alpha_2 (\mu - \Delta)$. We define the following distribution profiles $\nu^{(q)}$, for $q \in \{1, \ldots, K\}$:

$$
\nu_k^{(q)} = \begin{cases} 
\mathcal{N}(\hat{\mu}, \sigma^2) & \text{if } k = 1 \\
\mathcal{N}(\hat{\mu} + \Delta, \sigma^2) & \text{if } k = q \neq 1 \\
\mathcal{N}(\hat{\mu} - \Delta, \sigma^2) & \text{o.w.} 
\end{cases}
$$
That is, 

\[ \nu^{(1)} = \left( \begin{array}{c} \mathcal{N}(\mu, \sigma^2) \\ \mathcal{N}(\mu - \Delta, \sigma^2) \\ \vdots \\ \mathcal{N}(\mu - \Delta, \sigma^2) \end{array} \right), \quad \nu^{(2)} = \left( \begin{array}{c} \mathcal{N}(\mu, \sigma^2) \\ \mathcal{N}(\mu + \Delta, \sigma^2) \\ \vdots \\ \mathcal{N}(\mu + \Delta, \sigma^2) \end{array} \right), \quad \nu^{(3)} = \left( \begin{array}{c} \mathcal{N}(\mu, \sigma^2) \\ \mathcal{N}(\mu - \Delta, \sigma^2) \\ \vdots \\ \mathcal{N}(\mu - \Delta, \sigma^2) \end{array} \right), \ldots, \quad \nu^{(K)} = \left( \begin{array}{c} \mathcal{N}(\mu, \sigma^2) \\ \mathcal{N}(\mu - \Delta, \sigma^2) \\ \vdots \\ \mathcal{N}(\mu + \Delta, \sigma^2) \end{array} \right). \]

Accordingly, we define the distribution for auxiliary observations as follows:

\[ \nu^{(m,q)}_k = \begin{cases} \mathcal{N}(\frac{\mu}{\alpha_2}, 0) & \text{if } k = 1 \\ \mathcal{N}(\frac{\mu + \Delta}{\alpha_2}, 0) & \text{if } k = q \neq 1 \\ \mathcal{N}(\frac{\mu + \Delta}{\alpha_2}, 0) & \text{O.W.} \end{cases} \]

That is, 

\[ \nu^{(1)} = \left( \begin{array}{c} \mathcal{N}(\frac{\mu}{\alpha_2}, 0) \\ \mathcal{N}(\frac{\mu + \Delta}{\alpha_2}, 0) \\ \vdots \\ \mathcal{N}(\frac{\mu + \Delta}{\alpha_2}, 0) \end{array} \right), \quad \nu^{(2)} = \left( \begin{array}{c} \mathcal{N}(\frac{\mu}{\alpha_2}, 0) \\ \mathcal{N}(\frac{\mu + \Delta}{\alpha_2}, 0) \\ \vdots \\ \mathcal{N}(\frac{\mu + \Delta}{\alpha_2}, 0) \end{array} \right), \quad \nu^{(3)} = \left( \begin{array}{c} \mathcal{N}(\frac{\mu}{\alpha_2}, 0) \\ \mathcal{N}(\frac{\mu + \Delta}{\alpha_2}, 0) \\ \vdots \\ \mathcal{N}(\frac{\mu + \Delta}{\alpha_2}, 0) \end{array} \right), \ldots, \quad \nu^{(K)} = \left( \begin{array}{c} \mathcal{N}(\frac{\mu}{\alpha_2}, 0) \\ \mathcal{N}(\frac{\mu + \Delta}{\alpha_2}, 0) \\ \vdots \\ \mathcal{N}(\frac{\mu + \Delta}{\alpha_2}, 0) \end{array} \right). \]

One can see that each profile pair \((\nu^{(q)}, \nu^{(q)})\) is \(\mathcal{S}'\). We note that \(\nu^{(q)}\) is the same for \(q \in \{1, \ldots, K\}\), that is, roughly speaking, the auxiliary information does not carry any information regarding the true distribution profile \(\nu^{(q)}\). Given this observation, if we carry out similar steps as in the proof of Theorem 1, the desired result will be established.

A.8 Proof of Proposition 2

Define the event \(B_t := \{\forall k \in \mathcal{K}: \sum_{s=1}^t h_{k,s} \geq \lambda t / 2\}\). Using the Bernstein’s inequality in Lemma 6, one has \(\Pr \{B_t\} \geq 1 - K e^{\frac{\lambda t}{\lambda^2}}\). Also, define the event \(A_t := \{\forall k \in \mathcal{K}: |\tilde{Z}_{k,t} - \mu_k| \leq \sqrt{\frac{\theta^2}{2\lambda^2} \log t}\}\). Using the Chernoff-Hoeffding bound in Lemma 5, one obtains \(\Pr \{A_t \cap B_t\} \geq 1 - K e^{\frac{\lambda t}{\lambda^2}} - 2K t^{-2}\). Let \(t^* = \frac{27 \theta^2}{\lambda^2} \log \left( \frac{27 \theta^2}{\lambda^2} \right)\). On the event \(A_t \cap B_t\), for \(t \geq t^*\), we have

\[ \forall k \in \mathcal{K}: \frac{\Delta_k}{2} \leq \hat{\Delta}_{k,t} \leq 3 \Delta_k / 2. \]

(26)
Using the expression of $\tau_{k,t}$ from (16) we have the following inequality on the event $A_t \cap B_t$:

$$\tau_{k,t} \geq \exp \left( \frac{\lambda \Delta_k^2 t}{8c \sigma^2} \right). \quad (27)$$

One may decompose the regret as follows:

$$\sum_{k \in \mathcal{K}} \Delta_k \cdot \mathbb{E}_{(\nu,\nu')} \left[ \sum_{t=1}^T 1 \{ \pi_t = k \} \right] \leq \max_{k \in \mathcal{K}} \Delta_k t^* + \max_{k \in \mathcal{K}} \Delta_k \sum_{t=1}^T \mathbb{P} \{ A_t^c \cup B_t^c \}$$

$$+ \sum_{k \in \mathcal{K}} \Delta_k \cdot \left[ \sum_{t=t^*+1}^T \mathbb{P}_{(\nu,\nu')} \{ \pi_t = k, \text{ policy does exploration } | A_t \cap B_t \} \right]$$

$$+ \sum_{k \in \mathcal{K}} \Delta_k \cdot \left[ \sum_{t=t^*+1}^T \mathbb{P}_{(\nu,\nu')} \{ \pi_t = k, \text{ policy does exploitation } | A_t \cap B_t \} \right].$$

Using (26) and (27), one obtains

$$J_{k,1} \leq \sum_{t=1}^T \frac{4c \sigma^2}{\Delta_k^2} \exp \left( -\frac{\lambda \Delta_k^2 t}{8c \sigma^2} \right) \leq \frac{32c^2 \sigma^2 \sigma^2}{\lambda \Delta_k^4}.$$ 

Following the same argument as in Step 5 of the proof of Theorem 4, one can show that

$$J_{k,2} \leq 2 \sum_{t=1}^T 2 \exp \left( -\frac{\lambda \Delta_k^2 t}{16 \sigma^2} \right) \leq \frac{64 \sigma^2}{\Delta_k^2 \lambda}.$$ 

Finally, note that

$$\sum_{t=1}^T \mathbb{P} \{ A_t^c \cup B_t^c \} \leq 2K + 10K/\lambda,$$

which concludes the proof. ■

A.9 Proof of Proposition 3

Define the event $B_t := \left\{ \forall k \in \mathcal{K} : \sum_{s=1}^t h_{k,s} \geq \frac{\sigma_k^2}{\Delta_k} \log t \right\}$. Using the Bernstein’s inequality in Lemma 6, one has $\Pr \{ B_t \} \geq 1 - e^{-\frac{\sigma_k^2}{20 \Delta_k^2} \log t}$. Also, define the event

$$A_t := \left\{ |\hat{Z}_{k,t} - \mu_k| \leq \frac{\Delta_k}{4} \forall k \in \mathcal{K} \setminus \{ k^* \} \text{ and } |\hat{Z}_{k^*,t} - \mu_{k^*}| \leq \frac{\Delta_k}{4} \right\}.$$ 

Using the Chernoff-Hoeffding bound in Lemma 5, one obtains $\Pr \{ A_t \cap B_t \} \geq 1 - K e^{-\frac{\sigma_k^2}{20 \Delta_k^2} \log t} - 2K e^{-\frac{\sigma_k^2}{128} \log t}$. On the event $A_t \cap B_t$, we have

$$\forall k \in \mathcal{K} : \Delta_k/2 \leq \hat{\Delta}_{k,t} \leq 3\Delta_k/2.$$ 

(28)
Using the expression of $\tau_{k,t}$ from (16) we have the following inequality on the event $A_t \cap B_t$:

$$\tau_{k,t} \geq \exp\left(\frac{\kappa \log t}{16c}\right).$$

(29)

One may decompose the regret as follows:

$$\sum_{k \in K} \Delta_k \cdot \mathbb{P}_{(\nu,\nu')}\left[\sum_{t=1}^T \mathbbm{1}\{\pi_t = k\}\right] \leq \max_{k \in K} \sum_{t=1}^T \mathbb{P}\{A_t^c \cup B_t^c\}$$

$$+ \sum_{k \in K} \Delta_k \cdot \left[\sum_{t=1}^T \mathbb{P}_{(\nu,\nu')}\{\pi_t = k, \text{ policy does exploration } | A_t \cap B_t\}\right]$$

$$+ \sum_{k \in K} \Delta_k \cdot \left[\sum_{t=1}^T \mathbb{P}_{(\nu,\nu')}\{\pi_t = k, \text{ policy does exploitation } | A_t \cap B_t\}\right].$$

Using (28) and (29) we have

$$J_{k,1} \leq \sum_{t=1}^T \frac{4c\sigma^2}{\Delta_k^2} \exp\left(-\frac{\kappa \log t}{16c}\right) \leq \frac{16c^2\sigma^2}{\Delta_k^2(\kappa - 16c)}.$$

Following the same argument as in Step 5 of the proof of Theorem 4, one can show that

$$J_{k,2} \leq 2 \sum_{t=1}^T 2 \exp\left(-\frac{\kappa \Delta_k^2 \log t}{32\Delta^2}\right) \leq \frac{128\Delta^2}{(\kappa \Delta_k^2 - 32\Delta^2)}.$$

Finally, note that

$$\sum_{t=1}^T \mathbb{P}\{A_t^c \cup B_t^c\} \leq K \frac{20\Delta^2}{(\sigma^2 \kappa - 20\Delta^2)} + 2K \frac{128}{(\kappa - 128)},$$

which concludes the proof.

B Auxiliary lemmas

Lemma 3 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 for any measurable function $\Psi: \mathbb{X} \to \{0,1\}$, one has:

$$\mathbb{P}_{\rho_0}\{\Psi(X) = 1\} + \mathbb{P}_{\rho_1}\{\Psi(X) = 0\} \geq \frac{1}{2} \exp(-\text{KL}(\rho_0, \rho_1)).$$
Proof. Define \( A \) to be the event that \( \Psi(X) = 1 \). One has

\[
P_{\rho_0}\{\Psi(X) = 1\} + P_{\rho_1}\{\Psi(X) = 0\} = P_{\rho_0}\{A\} + P_{\rho_1}\{\bar{A}\} \geq \int \min\{d\rho_0, d\rho_1\} \geq \frac{1}{2} \exp(-\text{KL}(\rho_0, \rho_1)),
\]

where the last inequality follows from Tsybakov 2009, Lemma 2.6. ■

Lemma 4 Consider two profiles \( \nu \), and \( \hat{\nu} \). Denote by \( \nu_t(\hat{\nu}_t) \) the distribution of the observed rewards up to time \( t \) under \( \nu(\hat{\nu}) \) respectively. Let \( n_{k,t} \) be the number of times a sample from arm \( k \) has been observed up to time \( t \), that is

\[
n_{k,t} = h_{k,t} + \sum_{s=1}^{t-1} (h_{k,s} + 1 \{\pi_s = k\}).
\]

For any policy \( \pi \), we have

\[
\text{KL}(\nu_t, \hat{\nu}_t) = \sum_{k=1}^K E_\nu[n_{k,t}] \cdot \text{KL}(\nu_k, \hat{\nu}_k).
\]

Proof. The proof is a simple adaptation of the proof of Lemma 1 in Gerchinovitz and Lattimore (2016); hence, it is omitted. ■

Lemma 5 (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 \( E[X_t \mid X_1, \ldots, X_{t-1}] = \mu \). Let \( S_n = X_1 + \cdots + X_n \). Then for all \( a \geq 0 \)

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

Lemma 6 (Bernstein inequality) Let \( X_1, \ldots, X_n \) be random variables with range \([0,1]\) and

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

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

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

Lemma 7 Let \( X_1, \ldots, X_n \) be i.i.d. Bernoulli random variable with parameters \( p_1, \ldots, p_n \), respectively, then, for any \( \kappa > 0 \):

\[
E\left[e^{-\kappa \sum_{j=1}^n X_j}\right] \leq e^{-\kappa \sum_{j=1}^n p_j/10} + e^{-\kappa \sum_{j=1}^n p_j/2}.
\]
Proof. Define event $E$ to be the event that $\sum_{j=1}^{n} X_j < \sum_{j=1}^{n} p_j/2$. By Lemma 6, we have:

$$\mathbb{P}\{E\} \leq e^{-\sum_{j=1}^{n} p_j/10}.$$ 

By the law of total expectation:

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

\[ Q.E.D. \]

Lemma 8 For any $\kappa > 0$ and integers $n_1$ and $n_2$, we have $\sum_{j=n_1}^{n_2} e^{-\kappa j} \leq \frac{1}{\kappa}(e^{-\kappa(n_1-1)} - e^{\kappa n_2}).$

Proof. We can upper bound the summation using an integral as follows:

$$\sum_{j=n_1}^{n_2} e^{-\kappa j} \leq \int_{n_1-1}^{n_2} e^{-\kappa x} dx = \frac{1}{\kappa}(e^{-\kappa(n_1-1)} - e^{\kappa n_2}).$$

\[ Q.E.D. \]

Lemma 9 For any $t > 1/2$, and $n \in \{0, 1, 2, \ldots\}$, we have $\log\frac{t+n+1}{t} \leq \sum_{s=0}^{n} \frac{1}{t+s} \leq \log\frac{t+n+1/2}{t-1/2}$.

Proof. We first, show the lower through an integral:

$$\sum_{s=0}^{n} \frac{1}{t+s} \geq \int_{t}^{t+n+1} \frac{dx}{x} = \log\frac{t+n+1}{t}.$$ 

Now, we show the upper bound through a similar argument:

$$\sum_{s=0}^{n} \frac{1}{t+s} \leq \sum_{s=0}^{n} \int_{t+s-1/2}^{t+s+1/2} \frac{dx}{x} = \int_{t-1/2}^{t+n+1/2} \frac{dx}{x} = \log\frac{t+n+1/2}{t-1/2},$$

where (a) holds because we have $\frac{1}{t+s} \leq \int_{t+s-1/2}^{t+s+1/2} \frac{dx}{x}$ by Jensen’s inequality. \[ Q.E.D. \]

Lemma 10 (Pinsker’s inequality) Let $\rho_0, \rho_1$ be two probability distributions supported on some set $\mathcal{X}$, with $\rho_0$ absolutely continuous with respect to $\rho_1$ then,

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

51
where \( \| \rho_0 - \rho_1 \|_1 = \sup_{A \subset X} |\rho_0(A) - \rho_1(A)| \) is the variational distance between \( \rho_0 \) and \( \rho_1 \).

**Lemma 11** Let \( X_t \in \{0,1\}, 1 \leq t \leq T \), be zero-one random variables. Also, let \( h_t \geq 0 \), for \( 1 \leq t \leq T \), be some deterministic constants. Finally, define \( n_t = \sum_{s=1}^{t-1} (X_s + h_s) + h_t \), and assume \( \mathbb{E}[X_t | n_t] \leq p_{n_t+1} \) for some \( p \leq 1 \) then, we have

\[
\mathbb{E}[n_t] \leq \log_{\frac{1}{p}} \left( 1 + \sum_{s=1}^{t-1} p_s \sum_{\tau=s}^{t} h_{\tau} \right) + \sum_{s=1}^{t} h_s,
\]

for every \( 1 \leq t \leq T + 1 \), with \( h_{T+1} \) defined to be zero.

**Proof.** Consider the sequence \( p^{-n_t} \). For every \( t \),

\[
\mathbb{E}[p^{-n_t} | p^{-n_{t-1}}] = \mathbb{P}\{X_{t-1} = 1 | n_{t-1}\} \cdot p^{-(n_{t-1}+h_t+1)} + \mathbb{P}\{X_{t-1} = 0 | n_{t-1}\} \cdot p^{-(n_{t-1}+h_t)} \\
\leq p_{n_{t-1}+1} \cdot p^{-(n_{t-1}+h_t+1)} + p^{-(n_{t-1}+h_t)} \\
= p^{-h_t} \cdot (1 + p^{-n_{t-1}}).
\]

Using this inequality and a simple induction, one obtains

\[
\mathbb{E}[p^{-n_t}] \leq \sum_{s=1}^{t} p^{-\sum_{\tau=s}^{t} h_{\tau}},
\]

for every \( t \). Finally, by Jensen’s inequality, we have

\[
p^{-\mathbb{E}[n_t]} \leq \mathbb{E}[p^{-n_t}],
\]

which implies

\[
\mathbb{E}[n_t] \leq \log_{\frac{1}{p}} \left( \sum_{s=1}^{t} p^{-\sum_{\tau=s}^{t} h_{\tau}} \right) = \log_{\frac{1}{p}} \left( 1 + \sum_{s=1}^{t-1} p_s \sum_{\tau=s}^{t} h_{\tau} \right) + \sum_{s=1}^{t} h_s.
\]

**Lemma 12** Let \( X_t \in \{0,1\}, 1 \leq t \leq T \), be zero-one random variables, and define \( n_t = \sum_{s=1}^{t-1} X_s \), and assume \( \mathbb{E}[X_t | n_t] \leq q_{n_t+1}^\omega \) for some \( \omega \geq 1 \), \( q > 0 \), and \( p \leq 1 \) then, we have

\[
\mathbb{E}[n_t] \leq \sqrt{\log_{\frac{1}{p}} (q(T+1))},
\]

for every \( 1 \leq t \leq T + 1 \).
Proof. Consider the sequence $q^{-1} p^{-n_t^\omega}$. For every $t$,

$$
E \left[ q^{-1} p^{-n_t^\omega} \mid q^{-1} p^{-n_{t-1}^\omega} \right] = P \{ X_{t-1} = 1 \mid n_{t-1} \} \cdot q^{-1} p^{-(n_{t-1}+1)^\omega} + P \{ X_{t-1} = 0 \mid n_{t-1} \} \cdot q^{-1} p^{-n_{t-1}^\omega} \leq p^{(n_{t-1}+1)^\omega} \cdot p^{-(n_{t-1}+1)^\omega} + q^{-1} p^{-n_{t-1}^\omega} = 1 + q^{-1} p^{-n_{t-1}^\omega}.
$$

(34)

Using this inequality and a simple induction, one obtains

$$
E \left[ q^{-1} p^{-n_t^\omega} \right] \leq t + 1,
$$

for every $t$. Finally, by Jensen’s inequality, we have

$$
p^{-E \left[ n_t \right]^\omega} \leq E \left[ p^{-n_t^\omega} \right],
$$

which implies

$$
E \left[ n_t \right] \leq \sqrt{\log_\frac{1}{p} (q(t+1))}.
$$

\[\blacksquare\]

C Auxiliary analysis

C.1 Corollaries

Corollary 1 (Near optimality under stationary information flows) Let $\pi$ be Thompson sampling with auxiliary observations tuned by $c > 0$. If $h_{k,t}$’s are i.i.d. Bernoulli random variables with parameter $\lambda$ then, for every $T \geq 1$:

$$
E_H [\mathcal{R}_S^\omega (H, T)] \leq \left( \sum_{k \in K} \Delta_k \right) \left( \frac{18c \sigma^2}{\Delta^2} \log \left( \min \left\{ T+1, \frac{18c \sigma^2 + 10 \Delta^2}{\Delta^2 \lambda} \right\} \right) + C \left( 1 + \frac{1}{\Delta_k} \right) \right),
$$

for some constant $C$ independent of $T$. 

53
Proof.

\[
\mathbb{E}_H [\mathcal{R}_S^\pi (H, T)] \overset{(a)}{\leq} \mathbb{E}_H \left[ \sum_{k \in \mathcal{K}} \frac{18\sigma^2 \Delta_k}{\Delta^2} \log \left( \sum_{t=1}^{T} \exp \left( -\frac{\Delta^2}{18\sigma^2} \sum_{s=1}^{t} h_{k,s} \right) \right) + C \Delta_k \left( 1 + \frac{1}{\Delta_k^4} \right) \right]
\]

\[
\overset{(b)}{\leq} \sum_{k \in \mathcal{K}} \left[ \frac{18\sigma^2 \Delta_k}{\Delta^2} \log \left( \mathbb{E}_H \left[ \sum_{t=1}^{T} \exp \left( -\frac{\Delta^2}{18\sigma^2} \sum_{s=1}^{t} h_{k,s} \right) \right] \right) + C \Delta_k \left( 1 + \frac{1}{\Delta_k^4} \right) \right]
\]

\[
\overset{(c)}{\leq} \left( \sum_{k \in \mathcal{K}} \Delta_k \right) \cdot \left( \frac{18\sigma^2}{\Delta^2} \log \left( \frac{1}{10} + \frac{1}{\Delta^2 \lambda / 18\sigma^2} \right) + C \left( 1 + \frac{1}{\Delta_k^4} \right) \right),
\]

(35)

where: (a) holds by Theorem 2, (b) follows from Jensen’s inequality and the concavity of \( \log(\cdot) \), and (c) holds by Lemmas 7 and 8. Noting that \( \sum_{t=0}^{T} \exp \left( -\frac{\Delta^2}{18\sigma^2} \sum_{s=1}^{t} h_{k,s} \right) < T + 1 \), the result is established. 

**Corollary 2** (Near optimality under diminishing information flows) Let \( \pi \) be Thompson sampling with auxiliary observations tuned by \( c > 0 \). If \( h_{k,t} \)'s are random variables such that for some \( \kappa \in \mathbb{R}^+ \), \( \mathbb{E} \left[ \sum_{s=1}^{t} h_{k,s} \right] = |\frac{\sigma^2 \kappa}{2\Delta^2} \log t | \) for each arm \( k \in \mathcal{K} \) at each time step \( t \), then for every \( T \geq 1 \):

\[
\mathbb{E}_H [\mathcal{R}_S^\pi (H, T)] \leq \left( \sum_{k \in \mathcal{K}} \Delta_k \right) \left( \frac{18\sigma^2}{\Delta^2} \log \left( 2 + \frac{T^{1-\frac{\kappa}{72c}} - \frac{1}{1 - \frac{\kappa}{72c}} + \frac{T^{1-\frac{\kappa^2}{20\Delta^2}} - \frac{1}{1 - \frac{\kappa^2}{20\Delta^2}}}{\Delta^2 \lambda / 18\sigma^2} \right) + C \left( 1 + \frac{1}{\Delta_k^4} \right) \right),
\]

for some constant \( C \) independent of \( T \).

Proof.

\[
\mathbb{E}_H [\mathcal{R}_S^\pi (H, T)] \overset{(a)}{\leq} \mathbb{E}_H \left[ \sum_{k \in \mathcal{K}} \frac{18\sigma^2 \Delta_k}{\Delta^2} \log \left( \sum_{t=1}^{T} \exp \left( -\frac{\Delta^2}{18\sigma^2} \sum_{s=1}^{t} h_{k,s} \right) \right) + C \Delta_k \left( 1 + \frac{1}{\Delta_k^4} \right) \right]
\]

\[
\overset{(b)}{\leq} \sum_{k \in \mathcal{K}} \left[ \frac{18\sigma^2 \Delta_k}{\Delta^2} \log \left( \mathbb{E}_H \left[ \sum_{t=1}^{T} \exp \left( -\frac{\Delta^2}{18\sigma^2} \sum_{s=1}^{t} h_{k,s} \right) \right] \right) + C \Delta_k \left( 1 + \frac{1}{\Delta_k^4} \right) \right]
\]

\[
\overset{(c)}{\leq} \left( \sum_{k \in \mathcal{K}} \Delta_k \right) \cdot \left( \frac{18\sigma^2}{\Delta^2} \log \left( \sum_{t=1}^{T} \frac{t^{\frac{\kappa}{72c}}}{1 + \frac{t^{\frac{\kappa^2}{20\Delta^2}}}{\Delta^2 \lambda / 18\sigma^2}} \right) + C \left( 1 + \frac{1}{\Delta_k^4} \right) \right)
\]

\[
= \left( \sum_{k \in \mathcal{K}} \Delta_k \right) \cdot \left( \frac{18\sigma^2}{\Delta^2} \log \left( \sum_{t=1}^{T} t^{\frac{\kappa}{72c}} + t^{\frac{\kappa^2}{20\Delta^2}} \right) + C \left( 1 + \frac{1}{\Delta_k^4} \right) \right)
\]

\[
\leq \left( \sum_{k \in \mathcal{K}} \Delta_k \right) \cdot \left( \frac{18\sigma^2}{\Delta^2} \log \left( \sum_{t=1}^{T} t^{\frac{\kappa}{72c}} + t^{\frac{\kappa^2}{20\Delta^2}} \right) + C \left( 1 + \frac{1}{\Delta_k^4} \right) \right)
\]

\[
= \left( \sum_{k \in \mathcal{K}} \Delta_k \right) \cdot \left( \frac{18\sigma^2}{\Delta^2} \log \left( 2 + \int_{1}^{T} \left( t^{\frac{\kappa}{72c}} + t^{\frac{\kappa^2}{20\Delta^2}} \right) dt \right) + C \left( 1 + \frac{1}{\Delta_k^4} \right) \right),
\]

(36)
where: (a) holds by Theorem 2, (b) follows from Jensen’s inequality and the concavity of \( \log(\cdot) \), and (c) holds by Lemma 7. ■

### C.2 Proof of Subsection 3.1.1

We note that:

\[
E_H [R^\pi_S(H, T)] \geq \mathbb{E}_H \left[ \frac{\sigma^2(K - 1)}{4K} \log \left( \frac{\Delta^2}{\sigma^2 K} \sum_{t=1}^T \sum_{s=1}^T e^{-\frac{2\Delta^2}{\sigma^2} T \lambda} \right) \right]
\]

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

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

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

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 \( 1 - e^{-\frac{2\Delta^2}{\sigma^2} \lambda} \leq 2\Delta^2 \lambda / \hat{\sigma}^2 \). Now we consider the following cases:

1. If \( 2\Delta^2 T \lambda \leq \hat{\sigma}^2 / 2 \) then, by inequality \( 1 - e^{-x} \geq 2(1 - e^{-1/2})x \) for \( 0 \leq x \leq 1/2 \), and (39), we have

\[
E_H [R^\pi_S(H, T)] \geq \frac{\sigma^2(K - 1)}{4\Delta} \log \left( \frac{(1 - e^{-1/2}) \Delta^2 T}{\sigma^2 K} \right).
\]

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 (39), we have

\[
E_H [R^\pi_S(H, T)] \geq \frac{\sigma^2(K - 1)}{4\Delta} \log \left( \frac{1 - e^{-1/2}}{2K \lambda \sigma^2 / \hat{\sigma}^2} \right)
\]

C.3 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

\[
P\{\pi_t = k\} \leq P\{\bar{X}_{k,nk,t} > X_{k^*,nk,t}\} \leq P\left\{ \bar{X}_{k,nk,t} > \mu_k + \Delta_k / 2 \right\} + P\left\{ X_{k^*,nk,t} < \mu_{k^*} + \Delta_k / 2 \right\}
\]

We will upper bound \( P\left\{ \bar{X}_{k,nk,t} > \mu_k + \Delta_k / 2 \right\} \). \( P\left\{ \bar{X}_{k^*,nk,t} < \mu_{k^*} + \Delta_k / 2 \right\} \) can be upper bounded similarly.
\[ P \left\{ \bar{X}_{k,n,k,t} > \mu_k + \frac{\Delta_k}{2} \right\} \leq P \left\{ \bar{X}_{k,n,k,t} > \mu_k + \frac{\Delta_k}{2} \left| \sum_{s=1}^{t} h_{k,s} > \frac{\lambda t}{2} \right\} \right\} \leq P \left\{ \sum_{s=1}^{t} h_{k,s} \leq \frac{\lambda t}{2} \right\} \leq e^{-\frac{\Delta^2 \lambda}{16\sigma^2} + e^{-\frac{\lambda^2}{10}}}, \]

where (a) follows from Lemma 5 and Lemma 6. As a result the cumulative regret is upper bounded by

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

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

**C.4 Proof of Subsection 3.1.2.**

We note that:

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

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

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

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

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

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)). \( \square \)

**C.5 Analysis of the myopic policy under the setting of Example 3.1.2.**

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

\[
R_{(\nu, \nu')}^{\pi}(H, T) \leq \sum_{k \in K \setminus k^*} \Delta_k \cdot \sum_{t=1}^{T} t \left( e^{-\Delta_k^2 n_{k,t}} + e^{-\Delta_k^2 n_{k^*,t}} \right)
\]

\[
\leq \sum_{k \in K \setminus k^*} \Delta_k \cdot \sum_{t=1}^{T} t \left( \frac{-\Delta_k^2 \kappa}{16\Delta^2} + \frac{-\Delta_k^2 \kappa}{16\Delta^2} \right) \leq \sum_{k \in K \setminus k^*} \frac{32 \Delta_k^2}{\kappa \Delta_k}. \quad \square
\]

56
C.6 Reactive information flows

In our basic framework auxiliary information is assumed to be independent of the past decision trajectory of the policy. We next address some potential implications of endogenous information flows by considering a simple extension of our model in which information flows depend on past decisions. For the sake of concreteness, we assume that information flows are 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:

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

where $\tilde{n}_{k,t} = \sum_{s=1}^{t} \mathbb{1}\{\pi_s = k\}$ 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 $\alpha_{k,j} \geq 0$ captures the dependence of auxiliary observations of arm $k$ on the past selections of arm $j$. We assume that there exist non-negative values $\bar{\rho}$, $\rho$, $\bar{\alpha}$, and $\alpha$ such that $\rho \leq \rho_k \leq \bar{\rho}$ and $\alpha \leq \alpha_{k,j} \leq \bar{\alpha}$ for all arms $k, j \in \mathcal{K}$.

While the structure of (4) 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 is the order of the total number of times an arm is observed, $n_{k,t} = \tilde{n}_{k,t-1} + \sum_{s=1}^{t} \tilde{\sigma}_s^2 h_{k,t}$, 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; as this is the case when $\omega < 1$, we next 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, when $\gamma > 0$ and $\alpha > 0$), constant regret is achievable, and a myopic policy can guarantee rate-optimal performance. This observation is formalized by the following result.

**Proposition 4** 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 $\alpha > 0$ and $\gamma > 0$. Then, for any horizon length $T \geq 1$ and for any history-dependent information arrival matrix $H$ that satisfies (4), one has
\[ \mathcal{R}_S^\pi(H, T) \leq \sum_{k \in K \setminus \{k^*\}} \frac{C_{15} \cdot \Gamma(\frac{1}{\gamma})}{\hat{\gamma} \Delta_k^{\frac{1}{\gamma}-1}} + C_{16} \Delta_k, \]

where \( \Gamma(\cdot) \) is the gamma function, \( \hat{\gamma} = \min\{\gamma, 1\} \), and \( C_{15} \) is a positive constant that depends on \( \gamma, \alpha, \sigma, \) and \( \hat{\sigma} \), and \( C_{16} \) is an absolute constant.

**Proof.** Define \( \hat{\gamma} := \min\{1, \gamma\} \), and \( \hat{\alpha} := \min\{1, \alpha \cdot \frac{\sigma^2}{\hat{\sigma}^2}\} \). First, we will show that for any arm \( k \), and at any time step \( t \), we have

\[ n_{k,t} \geq \hat{\alpha}(t - 1)^{\hat{\gamma}}. \quad (41) \]

We will consider two cases. In the first case, assume \( \gamma \geq 1 \) then, one obtains

\[
n_{k,t} = \bar{n}_{k,t-1} + \sum_{s=1}^{t} \frac{\sigma^2}{\hat{\sigma}^2} h_{k,s} \geq \bar{n}_{k,t-1} + \sum_{j \in K \setminus \{k\}} \alpha \cdot \frac{\sigma^2}{\hat{\sigma}^2} \bar{n}_{j,t-1}^{\gamma} \]

\[
\geq \bar{n}_{k,t-1} + \sum_{j \in K \setminus \{k\}} \alpha \cdot \frac{\sigma^2}{\hat{\sigma}^2} \bar{n}_{j,t-1}^{\gamma} \]

\[
\geq \min\{1, \alpha \cdot \frac{\sigma^2}{\hat{\sigma}^2}\} \sum_{j=1}^{K} \bar{n}_{j,t-1}^{\gamma} \geq \hat{\alpha}(t - 1)^{\hat{\gamma}},
\]

In the second case, assume \( \gamma < 1 \). One obtains

\[
n_{k,t} \geq \bar{n}_{k,t-1} + \sum_{j \in K \setminus \{k\}} \alpha \cdot \frac{\sigma^2}{\hat{\sigma}^2} \bar{n}_{j,t-1}^{\gamma} \geq \bar{n}_{k,t-1}^{\gamma} + \sum_{j \in K \setminus \{k\}} \alpha \cdot \frac{\sigma^2}{\hat{\sigma}^2} \bar{n}_{j,t-1}^{\gamma} \]

\[
\geq \min\{1, \alpha \cdot \frac{\sigma^2}{\hat{\sigma}^2}\} \sum_{j=1}^{K} \bar{n}_{j,t-1}^{\gamma} \geq \hat{\alpha}(t - 1)^{\hat{\gamma}},
\]

where the last step follows from the fact that \( x^\gamma + y^\gamma \geq (x + y)^\gamma \) for \( x, y \geq 0 \), and \( \gamma < 1 \). Hence, (41) is established. In order to upper bound the regret, we will use the usual decomposition

\[
\mathcal{R}_S^\pi(H, T) = \sum_{t=1}^{T} \sum_{k \in K \setminus \{k^*\}} \Delta_k \mathbb{P}^\pi_{(\nu, \nu')} \{ \pi_t = k \}.
\]

Fix \( k \neq k^* \). We upper bound each summand as follows:

\[
\mathbb{P}^\pi_{(\nu, \nu')} \{ \pi_t = k \} \leq \mathbb{P}^\pi_{(\nu, \nu')} \{ \bar{X}_{k,n_{k,t}} > \bar{X}_{k^*,n_{k^*,t}} \} \leq \mathbb{P}^\pi_{(\nu, \nu')} \left\{ \bar{X}_{k,n_{k,t}} > \mu_k + \frac{\Delta_k}{2} \right\} + \mathbb{P}^\pi_{(\nu, \nu')} \left\{ \bar{X}_{k^*,n_{k^*,t}} < \mu_{k^*} - \frac{\Delta_k}{2} \right\}.
\]

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

\[
P_\pi(\nu, \nu') \left\{ \bar{X}_{k,n_k,t} > \mu_k + \frac{\Delta_k}{2} \right\} \leq \mathbb{E}_P(\nu, \nu') \left[ \exp \left( \frac{-\Delta_k^2}{8\sigma^2 n_k,t} \right) \right] \leq \exp \left( \frac{-\Delta_k^2}{8\sigma^2 \tilde{\alpha}(t - 1)\tilde{\gamma}} \right),
\]

where (a), and (b) hold by Lemma 5, and (41), respectively. Finally, note that

\[
\sum_{t=1}^{T} P_\pi(\nu, \nu') \{ \pi_t = k \} \leq \sum_{t=1}^{T} 2 \exp \left( \frac{-\Delta_k^2}{8\sigma^2 \tilde{\alpha}(t - 1)\tilde{\gamma}} \right) \leq 2 + \frac{\Gamma(\frac{1}{\tilde{\gamma}})}{\tilde{\gamma}} \left( \frac{8\sigma^2}{\tilde{\alpha} \Delta_k^2} \right)^{1/\tilde{\gamma}},
\]

where \( \Gamma(\cdot) \) is the gamma function, and the last inequality follows from \( \sum_{t=1}^{\infty} \exp(-\kappa t^{\xi}) \leq \int_{0}^{\infty} \exp(-\kappa t^{\xi}) dt = \frac{\Gamma(\frac{1}{\xi})}{\kappa^{\frac{1}{\xi}}} \) for \( \kappa > 0 \) and \( 0 < \xi \leq 1 \).

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. While we evaluate the performance of our approach using the \( \epsilon_t \)-greedy with adaptive exploration policy, similar guarantees can also be established using the rest of the policies provided in this paper.

**Proposition 5 (Near Optimality under decoupled endogenous information flows)** Assume that \( \gamma = 0 \) and \( \omega \geq 1 \). For any \( T \geq 1 \) and history-dependent information arrival matrix \( \mathbf{H} \) that satisfies (4). Let \( \pi \) be an \( \epsilon_t \)-greedy policy with adaptive exploration, tuned by \( c > \max \left\{ 16, \frac{10\Delta^2}{\sigma^2} \right\} \). Then, for any \( T \geq \left\lceil \frac{cK\sigma^2}{\Delta^2} \right\rceil \):

\[
R^\pi_S(\mathbf{H},T) \leq C_{17} \cdot \left( \log \frac{T}{\Delta^2} \right)^{1/\omega} \cdot \sum_{k \in K \setminus \{k^*\}} \frac{\Delta_k}{\rho_k \Delta^2} + \sum_{k \in K \setminus \{k^*\}} C_{18} \Delta_k \left( 1 + \frac{1}{\Delta^2} \right),
\]

where \( C_{17}, \) and \( C_{18} \) are positive constants that depend on \( \sigma, \hat{\sigma}, \) and \( c \).

Proposition 5 introduces improved upper bound under the class of reactive information flows defined by (4) with \( \gamma = 0 \) and \( \omega \geq 1 \). For example, Proposition 5 implies that under this class, whenever mean rewards are separated (that is, \( \Delta_k \) is independent of \( T \) for each \( k \)), the best achievable regret is of order \( (\log T)^{1/\omega} \). We note that in proving Proposition 5 we assume no prior knowledge of either the class of information arrival processes given by (4), or the parametric values under this structure.

**Proof.** Let \( t^* = \left\lceil \frac{Kc\sigma^2}{\Delta^2} \right\rceil \), and fix \( T \geq t^*, K \geq 2, \) and profile \( (\nu, \nu') \in S \).

**Step 1 (Decomposing the regret).** We decompose the regret of a suboptimal arm \( k \neq k^* \) as follows:
\[
\Delta_k \cdot \mathbb{E}_{(\nu, \nu')}^\pi \left[ \sum_{t=1}^T \mathbb{1} \{ \pi_t = k \} \right] = \Delta_k \cdot \mathbb{E}_{(\nu, \nu')}^\pi \left[ \sum_{t=1}^T \mathbb{1} \{ \pi_t = k, \text{ policy does exploration} \} \right] \\
+ \Delta_k \cdot \mathbb{E}_{(\nu, \nu')}^\pi \left[ \sum_{t=1}^T \mathbb{1} \{ \pi_t = k, \text{ policy does exploitation} \} \right].
\]

(42)

The component \( J_{k,1} \) is the expected number of times arm \( k \) is pulled due to exploration, while 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 for \( t \geq t^* \):

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

(43)

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

\[
\hat{n}_{k,t} := \sum_{s=1}^{t} \mathbb{1} \{ \pi_s = k, \text{ policy does exploration} \}.
\]

Note that for \( t \geq t^* \),

\[
\mathbb{E} \left[ \mathbb{1} \{ \pi_t = k, \text{ policy does exploration} \} \mid \hat{n}_{k,t-1} \right] \leq \frac{c \sigma^2}{\Delta^2 \tau_{k,t}} \leq \frac{c \sigma^2}{\Delta^2} \exp \left( -\frac{\Delta^2}{c \sigma^2} \sum_{s=1}^{t} h_{k,s} \right) \leq \frac{c \sigma^2}{\Delta^2} \exp \left( -\frac{\Delta^2 \rho}{c \sigma^2} (\hat{n}_{k,t-1} + 1)^\omega \right).
\]

This inequality implies that we can apply Lemma 12 to the sequence \( \hat{n}_{k,t} \) to drive the following upper bound:

\[
\mathbb{E} [\hat{n}_{k,T}] \leq \sqrt{\frac{c \sigma^2}{\Delta^2} \log \left( \frac{c \sigma^2 (T + 1)}{\Delta^2} \right)}.
\]

Finally, one obtains

\[
J_{k,1} = \mathbb{E} [\hat{n}_{k,T}] \leq \sqrt{\frac{c \sigma^2}{\Delta^2} \log \left( \frac{c \sigma^2 (T + 1)}{\Delta^2} \right)}.
\]
**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 $t \geq t^*$. Fix $t \geq t^*$. Then, $n_{k,t}$ is a summation of independent Bernoulli r.v.’s. One has

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

(44)

Let $\bar{M}_t = \sum_{s=1}^{t} h_{k,s}$ and $\bar{M} = \sum_{s=1}^{t^*} h_{k,s}$. For $0 \leq m \leq \bar{M}_t$, 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 \left| \sum_{s=1}^{q} h_{k,s} = m \right. \right\} \right) 
    & \text{if } 0 \leq m \leq \bar{M}_t \\
    t - 1 & \text{if } m = \bar{M}_t + 1
\end{cases}
\]

One obtains

\[
\mathbb{E} \left[ \sum_{s=1}^{t} \frac{\sigma^2}{\sigma^2} h_{k,s} + \sum_{s=t^*}^{t-1} \frac{c\sigma^2}{\Delta^2} \tau_{k,s} \right] \geq \mathbb{E} \left[ \sum_{s=1}^{t^*} \frac{\sigma^2}{\sigma^2} h_{k,s} + \sum_{m=0}^{\bar{M}_t+1} \left( \frac{\sigma^2}{\Delta^2} \sum_{s=0}^{t_{m+1,t}-t_{m,t}-1} \frac{c\sigma^2}{\Delta^2 (\tau_{k,t_{m+1,t}} + s)} \right) \right] - \frac{\sigma^2}{\bar{M}_t} ^2
\]

(a)

\[
= \mathbb{E} \left[ \sum_{s=1}^{t^*} \frac{\sigma^2}{\sigma^2} h_{k,s} + \sum_{m=0}^{\bar{M}_t+1} \left( \frac{\sigma^2}{\Delta^2} \sum_{s=0}^{t_{m+1,t}-t_{m,t}-1} \frac{c\sigma^2}{\Delta^2 (\tau_{k,t_{m+1,t}} + s)} \right) \right] - \frac{\sigma^2}{\bar{M}_t} ^2
\]

= \mathbb{E} \left[ \sum_{s=1}^{t^*} \frac{\sigma^2}{\sigma^2} h_{k,s} + \sum_{m=0}^{\bar{M}_t+1} \left( \frac{\sigma^2}{\Delta^2} \sum_{s=0}^{t_{m+1,t}-t_{m,t}-1} \frac{c\sigma^2}{\Delta^2 (\tau_{k,t_{m+1,t}} + s)} \right) \right] - \frac{\sigma^2}{\bar{M}_t} ^2
\]

\[
\geq \mathbb{E} \left[ \frac{c\sigma^2}{\Delta^2} \log \left( \frac{\tau_{k,t-1}}{t^*} \right) \right] - \frac{\sigma^2}{\bar{M}_t} ^2
\]

\[
\geq \mathbb{E} \left[ \frac{c\sigma^2}{\Delta^2} \log \left( \frac{t - 1}{t^*} \right) \right] - \frac{\sigma^2}{\bar{M}_t} ^2
\]

(45)

where (a) follows from Lemma 9. Putting together (44) and (45), one obtains

\[
\mathbb{E} [n_{k,t}] \geq \frac{c\sigma^2}{\Delta^2} \log \left( \frac{t - 1}{t^*} \right) - \frac{\sigma^2}{\bar{M}_t} ^2 := x_{t}.
\]

(46)

By Bernstein’s inequality in Lemma 6 we have

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

(47)

where the last inequality follows from (43).
Step 5 (Analysis of $J_{k,2}$). We note that

$$\sum_{k \in K \setminus k^*} J_{k,2} \leq t^* - 1 + \sum_{k \in K \setminus k^*} \sum_{t=1}^{T} \mathbb{P}\left\{ \bar{X}_{k,n_k,t} > \bar{X}_{k^*,n_{k^*},t} \right\}. \quad (48)$$

We upper bound each summand as follows:

$$\mathbb{P}\left\{ \bar{X}_{k,n_k,t} > \bar{X}_{k^*,n_{k^*},t} \right\} \leq \mathbb{P}\left\{ \bar{X}_{k,n_k,t} > \mu_k + \frac{\Delta_k}{2} \right\} + \mathbb{P}\left\{ \bar{X}_{k^*,n_{k^*},t} < \mu_k^* + \frac{\Delta_k}{2} \right\}. \quad (49)$$

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

$$\mathbb{P}\left\{ \bar{X}_{k,n_k,t} > \mu_k + \frac{\Delta_k}{2} \right\} \overset{(a)}{\leq} \mathbb{P}\left\{ \bar{X}_{k,n_k,t} > \frac{x_t}{2} \right\} + \mathbb{P}\left\{ n_{k,t} \leq \frac{x_t}{2} \right\} \overset{(b)}{\leq} \exp\left( -\frac{\Delta_k^2 x_t}{16\sigma^2} \right) + \exp\left( -\frac{-\sigma^2}{10K} \cdot \left( \frac{t^*}{t-1} \right)^{\frac{-\sigma^2}{10\Delta^2}} \right) \leq \exp\left( \frac{\Delta_k^2}{16K\sigma^2} \right) \cdot \exp\left( \frac{-c}{16} \log \left( \frac{t-1}{t^*} \right) \right) + \exp\left( -\frac{\sigma^2}{10K} \cdot \left( \frac{t^*}{t-1} \right)^{\frac{-\sigma^2}{10\Delta^2}} \right) \leq \exp\left( \frac{\Delta_k^2}{16K\sigma^2} \right) \cdot \left( \frac{t^*}{t-1} \right)^{\frac{-c}{16}} + \exp\left( -\frac{\sigma^2}{10K} \cdot \left( \frac{t^*}{t-1} \right)^{\frac{-\sigma^2}{10\Delta^2}} \right), \quad (50)$$

where, (a) holds by the conditional probability definition, (b) follows from Lemma 5 together with (47), and (c) holds by (43). Putting together (42), (48), (49), and (50), the result is established.

\section{D Numerical analysis}

\subsection{D.1 Comparing empirical performance of various policies}

\textbf{Setup}. We simulate the performance of different policies using MAB instances with three arms. The reported results correspond to rewards that are Gaussian with means $\mu = \begin{pmatrix} 0.7 \\ 0.5 \\ 0.5 \end{pmatrix}$ and standard deviation $\sigma = 0.5$, but we note that results are very robust with respect to the reward distributions. For simplicity we drew auxiliary information from the same distributions as the rewards. We considered two information arrival processes: stationary information flows with $h_{k,t}$’s that are i.i.d. Bernoulli random variables with mean $\lambda$, where we considered values of $\lambda \in \{500/T, 100/T, 10/T\}$; and diminishing information flows where $h_{k,t}$’s are i.i.d. Bernoulli random variables with mean $\kappa'$, where we considered values of $\kappa' \in \{4, 2, 1\}$ for each time period $t$ and for each arm $k$.

In this part we experimented with three policies: $\epsilon_t$-greedy, UCB1, and Thompson sampling. We considered several variants of each policy: (i) a version that ignores auxiliary information arrivals (EG,
UCB1, TS); (ii) version that updates empirical means and cumulative number of observations upon information arrivals but uses a standard time index (nEG, nUCB1); and (iii) version that updates empirical means and cumulative number of observations upon information arrivals and uses virtual time indexes (aEG, aUCB, aTS1). The difference between these versions is essentially in manner in which empirical means and reward observation counters are computed, as well as in the type of time index that is used. Namely, for the original versions:

\[
\bar{X}_{k,n,t} := \frac{\sum_{s=1}^{t} \frac{1}{\sigma^2} \mathbb{1}\{\pi_s = k\} X_{k,s}}{\sum_{s=1}^{t} \frac{1}{\sigma^2} \mathbb{1}\{\pi_s = k\}}, \quad n_{k,t} := \sum_{s=1}^{t} \mathbb{1}\{\pi_s = k\}, \quad \tau_{k,t} = \tau_{k,t-1} + 1,
\]

for the versions that account for auxiliary information while using standard time index:

\[
\bar{X}_{k,n,t} := \frac{\sum_{s=1}^{t} \frac{1}{\sigma^2} \mathbb{1}\{\pi_s = k\} X_{k,s} + \sum_{s=1}^{t} \frac{1}{\sigma^2} Z_{k,s}}{\sum_{s=1}^{t} \frac{1}{\sigma^2} \mathbb{1}\{\pi_s = k\} + \sum_{s=1}^{t} \frac{1}{\sigma^2} h_{k,s}}, \quad n_{k,t} := \sum_{s=1}^{t} \mathbb{1}\{\pi_s = k\} + \sum_{s=1}^{t} \frac{\sigma^2}{\sigma^2} h_{k,s}, \quad \tau_{k,t} = \tau_{k,t-1} + 1,
\]

and for the versions with virtual time indexes:

\[
\bar{X}_{k,n,t} := \frac{\sum_{s=1}^{t} \frac{1}{\epsilon^2} \mathbb{1}\{\pi_s = k\} X_{k,s} + \sum_{s=1}^{t} \frac{1}{\epsilon^2} Z_{k,s}}{\sum_{s=1}^{t} \frac{1}{\epsilon^2} \mathbb{1}\{\pi_s = k\} + \sum_{s=1}^{t} \frac{1}{\epsilon^2} h_{k,s}}, \quad n_{k,t} := \sum_{s=1}^{t} \mathbb{1}\{\pi_s = k\} + \sum_{s=1}^{t} \frac{\epsilon^2}{\epsilon^2} h_{k,s}, \quad \tau_{k,t} = \tau_{k,t-1} + 1 \cdot \exp\left(\frac{h_{k,t} \Delta^2}{c_\pi \sigma^2}\right),
\]

where \(c_\pi\) depends on the policy type \(\pi\). The nUCB1 policy is the UCB1 policy with standard time index that is detailed in \[4.2\] and for which we have already provided near-optimal regret bound. The aUCB1 policy applies the virtual time indexes method, and can be described as follows. Define \(n_{k,t}\) and \(X_{n_{k,t},t}\) as in \(2\), and consider the following adaptation of UCB1 policy (Auer et al. 2002).

**UCB1 with adaptive exploration.** Inputs: a constant \(c\).

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 \(h_t\) and \(Z_t\)
   
   (b) Update the virtual times: \(\tau_{k,t} = (\tau_{k,t-1} + 1) \cdot \exp\left(\frac{h_{k,t} \Delta^2}{4c_\pi \sigma^2}\right)\) for all \(k \in \mathcal{K}\)
   
   (c) Select the arm

   \[
   \pi_t = \begin{cases} 
   t & \text{if } t \leq K \\
   \arg\max_{k \in \mathcal{K}} \left\{ \bar{X}_{k,n_{k,t}} + \sqrt{\frac{c_\pi \log \tau_{k,t}}{n_{k,t}}} \right\} & \text{if } t > K
   \end{cases}
   \]

   (d) Receive and observe a reward \(X_{\pi_t,t}\)
We next establish that UCB1 with virtual time indexes guarantees rate-optimal performance as well.

**Theorem 6** Let $\pi$ be a UCB1 with adaptive exploration policy, tuned by $c > 2$. Then, for any $T \geq 1$, $K \geq 2$, and auxiliary information arrival matrix $H$:

$$R_{\pi}^S(H, T) \leq \sum_{k \in K} \frac{C_{19}}{\Delta_k^2} \log \left( \sum_{t=0}^{T} \exp \left( -\frac{\Delta_t^2}{C_{19}} \sum_{s=1}^{t} h_{k,s} \right) \right) + C_{20} \Delta_k,$$

where $C_{19}$, and $C_{20}$ are positive constants that depend only on $\sigma$ and $\hat{\sigma}$.

The proof of Theorem 6 adapts the proof of Theorem 3, where now $l_{k,t} = \frac{4c\sigma^2 \log(\tau_{k,t})}{\Delta_k^2}$.

We measured the average empirical regret $\sum_{t=1}^{T} (\mu^* - \mu_{\pi_t})$, that is, the average difference between the performance of the best arm and the performance of the policy, over a decision horizon of $T = 10^4$ periods (performance over a short problem horizon is discussed in §D.4). Averaging over 400 repetitions the outcome approximates the expected regret. To determine the tuning parameter $c$ used by the policies we simulated their original versions without additional information flows for $c = 0.1 \times m$, where we considered values of $m \in \{1, 2, \ldots, 20\}$. The results presented below include the tuning parameter that resulted in the lowest cumulative empirical regret over the entire time horizon: $c = 1.0$ for $\epsilon_t$-greedy, $c = 1.0$ for UCB1, and $c = 0.5$ for Thompson sampling (robustness with respect to this tuning parameter is discussed in §D.2).

**Results and discussion.** Plots comparing the empirical regret of the different algorithms for stationary and diminishing information flows appear in Figures 5 and 6, respectively. For UCB policy, we observe that when the amount of auxiliary information on all arms is very large, UCB with a standard time index tends to have slightly smaller regret, whereas if there is less auxiliary information a UCB algorithm that uses virtual time indexes has a better performance. This stems from the fact that if we do not update the virtual time index in this policy, it tends to explore less after receiving auxiliary information. More precisely, in the naive UCB1 policy, the upper confidence bound shrinks faster than its adaptive version, which results in less exploration. For the $\epsilon_t$-greedy policy, as expected, using virtual time indexes always outperforms using a standard time index.

In the case of stationary information flows with large $\lambda$, the cumulative regret is bounded by a constant and does not grow after some point, and when $\lambda$ is smaller we transition back to the logarithmic regret, as discussed in §3.1.1. The impact of the timing of information arrivals on the achievable performance can be viewed by comparing the plots that correspond to the stationary and the diminishing information flows. The expected total number of information arrivals in the first, second, and third columns with diminishing information flows is less than or equal to that of the first, second, and third column with stationary information flows, and the cumulative regrets for both cases are very close to one another.
Figure 5: Cumulative regret trajectory of the policies studied in the paper under stationary information flow at a constant rate $\lambda$. Left column: $\lambda = \frac{500}{T}$; Middle column: $\lambda = \frac{100}{T}$; Right column: $\lambda = \frac{10}{T}$
Figure 6: Cumulative regret trajectory of the policies studied in the paper under diminishing information flow at rate $\frac{\kappa}{T}$ at each time step $t$. Left column: $\kappa' = 4$; Middle column: $\kappa' = 2$; Right column: $\kappa' = 1$

D.2 Misspecification of the tuning parameter $c$

**Setup.** We use the setup discussed in §D.1 to study the effect of using different values for the tuning parameter $c$. The plots here detail results for parametric values of $c = \{0.4, 1.0, 1.6\}$ for $\epsilon_t$-greedy, $c \in \{0.4, 1.0, 1.6\}$ for UCB1, and $c = \{0.1, 0.5, 0.7\}$ for Thompson sampling, and using four different information arrival instances: stationary with $\lambda \in \{500/T, 10/T\}$ and diminishing with $\kappa' \in \{4.0, 1.0\}$.

**Results and discussion.** Plots comparing the empirical regret accumulation of the policies with adaptive exploration with different tuning parameters $c$ for the stationary and diminishing information flows appear in Figure 7.

For the policies $\epsilon_t$-greedy, and Thompson sampling, one may observe that when there is a lot of
auxiliary information (the first and third columns) lower values of $c$ lead to better performance. However, when there is less auxiliary information (the second and fourth columns) very small values of $c$ may result in high regret. The reason for this is that when $c$ is small these policies tend to explore less, hence, in the presence of a lot of auxiliary information they incur little regret but if auxiliary information is rare these policies under-explores and incur a lot of regret. We observe a similar effect for UCB1 as well.

Figure 7: The effect of the tuning parameter $c$ on the performance of different policies under different information flows. First column on the left: stationary information flow at a constant rate $\lambda = \frac{500}{T}$. Second column: stationary information flow at a constant rate $\lambda = \frac{100}{T}$. Third column: diminishing information flow at rate $\kappa' t$ at each time step $t$ for $\kappa' = 8.0$. Fourth column: diminishing information flow at rate $\kappa' t$ at each time step $t$ for $\kappa' = 1.0$. (For description of the policies see §D.1)

D.3 Misspecification of the minimum gap $\Delta$ for $\epsilon_t$-greedy with adaptive exploration

Setup. We use the setup from §D.1 to study the effect of misspecifying the minimum gap $\Delta$ for $\epsilon_t$-greedy with virtual time indexes. We experiment with values of $\hat{\Delta}$ that are higher than, lower than, and equal to the true value of $\Delta = 0.2$. The representative results below detail the empirical regret incurred by using $\hat{\Delta} \in \{0.05, 0.2, 0.35\}$, in different information flows instances: stationary flows with $\lambda \in \{\frac{500}{T}, \frac{10}{T}\}$ and diminishing flows with $\kappa' \in \{8.0, 1.0\}$. We used the tuning parameter $c = 1.0$ in all cases.
Results and discussion. Plots comparing the regret accumulation with different specified gaps $\hat{\Delta}$ appear in Figure 8. In all cases, a conservative choice of the gap $\hat{\Delta}$ (that is, $\hat{\Delta} = 0.05$) results in over-exploration and additional regret compared with the case that the exact value of the gap is used. On the other hand, a lower value of $\hat{\Delta}$ (that is, $\hat{\Delta} = 0.4$) results in lower regret when the amount of information is high. The reason is that a higher value of $\hat{\Delta}$ leads to less exploration, which, when compensated by the presence of many auxiliary observations, leads to better performance overall. However, a higher value of $\hat{\Delta}$ can result in high regret when the amount of auxiliary information is low.

![Figure 8](image-url)

Figure 8: The effect of the gap parameter $\hat{\Delta}$ on the performance of $\epsilon_t$-greedy with adaptive exploration under different information flows. From left to right: stationary information flow at a constant rate $\lambda = \frac{500}{T}$; stationary information flow at a constant rate $\lambda = \frac{100}{T}$; diminishing information flow at rate $\frac{\kappa'}{T}$ at each time step $t$ for $\kappa' = 8.0$; diminishing information flow at rate $\frac{\kappa'}{T}$ at each time step $t$ for $\kappa' = 1.0$. (For description of the policies see §D.1.)

D.4 Short problem horizon

One can benefit from utilizing information flows also when the time horizon is small. Here we present numerical results that consider a smaller value of $T = 100$ with quantile information presented, i.e., box plots. We consider the same setting as in §D.1. We have simulated each policy over the time horizon length $T = 100$ for 2,000 replications. We considered stationary information flows with $h_{k,t}$'s being i.i.d. Bernoulli random variables with mean $\lambda$, where $\lambda \in \{20/T, 10/T, 5/T\}$, and diminishing information flows where $h_{k,t}$'s are i.i.d. Bernoulli random variables with mean $\frac{\kappa'}{T}$, where $\kappa' \in \{5, 2.5, 1.25\}$ for each time period $t$ and for each arm $k$. The box plots of the total regret can be found in Figures 9 and 10.

One may observe that leveraging auxiliary information improves the performance of all policies. For the $\epsilon_t$-greedy policy, the version with virtual time indexes performs the best, as expected. For the UCB1 policy the performance using a standard time index is almost the same as in using virtual time indexes. For Thompson sampling, including the auxiliary observations in updating the posterior is very beneficial.
Figure 9: The box plot of the total regret in short time horizon regime, $T = 100$, for stationary information flow at a constant rate $\lambda$. *Left column*: $\lambda = \frac{20}{T}$; *middle column*: $\lambda = \frac{10}{T}$; *right column*: $\lambda = \frac{5}{T}$. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the first and third quantiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the red ‘+’ symbol. (For description of the policies see §D.1.)
D.5 Utilizing auxiliary information flows for estimating the gaps $\Delta_k$

As it was discussed in §7.2, one might estimate the gaps $\Delta_k$ using auxiliary observations and use these estimates for tuning the MAB policy. To investigate this approach numerically, we analyze the performance of $\epsilon_t$-greedy with virtual time indexes and estimated gaps, detailed in §7.2 (denoted here by aEGest). We considered the same setup as in §D.1 and simulated this policy, along with the other variants of $\epsilon_t$-greedy that make use of the exact value of the gap (denoted by EGacc, nEGacc, and aEGacc) over the time horizon $T = 10,000$ for 400 replications with the tuning parameter $c = 1.0$ for all these policies. We considered stationary information flows with $h_{k,t}$'s being i.i.d. Bernoulli random variables with mean $\lambda$, where $\lambda \in \{2000/T, 1000/T, 500/T, 100/T\}$; and diminishing information flows where $h_{k,t}$'s are i.i.d. Bernoulli random variables with mean $\kappa' / T$, where $\kappa' \in \{16, 8, 4\}$.
Results and discussion. Representative examples of plots comparing the performance of $\epsilon_t$-greedy with adaptive exploration and estimated gaps and other $\epsilon_t$-greedy variants that make use of the exact value of the gap parameters appear in Figures 11 and 12. One may observe that in the presence of sufficient auxiliary observation $\epsilon_t$-greedy with adaptive exploration and estimated gaps achieves constant regret, and in some cases it outperforms the naive adaptation of $\epsilon_t$-greedy with an accurate gap parameter.

![Figure 11: Cumulative regret trajectory of different variants of $\epsilon_t$-greedy policy with accurate and estimated gap under stationary information flow at a constant rate $\lambda$. First column on the left: $\lambda = \frac{1000}{T}$; Second column: $\lambda = \frac{500}{T}$. (For description of the policies see §D.1.](image1)

![Figure 12: Cumulative regret trajectory of different variants of $\epsilon_t$-greedy policy with accurate and estimated gap under diminishing information flow at rate $\kappa'$ at each time step $t$. First column on the left: $\kappa' = 16$; Second column: $\kappa' = 8$; Third column: $\kappa' = 4$. (For description of the policies see §D.1.)](image2)

E Setup of analysis with real data

For each article we used the following setting. We adopted the times the article was recommended from first position (highest in a list of 5 recommended links) as the sequence of decision epochs for experimenting with that article. We focused on articles that were frequently recommended, and for consistency set $T = 2,000$ for each, ignoring subsequent epochs. For a fixed article $a$ and day $d$, denote by $\text{CTR}_{a,d}$ the fraction of readers that clicked the article’s title during day $d$ when this article was recommended to them in first position (click-through rate), and by $\text{CVR}_{a,d}^{\text{recom}}$ the fraction of readers that clicked on a recommendation hosted by the article during day $d$ after arriving to that article
through content recommendation (convergence rate). For \( k \in \{0, 1\} \) we define the (revenue-normalized) one-step lookahead expected recommendation value by \( \mu_k = \text{CTR}_{a,d,k}(1 + \text{CVR}_{\text{recom}}_{a,d,k}) \). When arm \( k \) is pulled at time \( t \), the agent observes the reward \( X_{k,t}^{(1)} (1 + X_{k,t}^{(2)}) \), where \( X_{k,t}^{(1)} \sim \text{Ber}(\text{CTR}_{a,d,k}) \) and \( X_{k,t}^{(2)} \sim \text{Ber}(\text{CVR}_{\text{recom}}_{a,d,k}) \).

The unknown arm has a conversion rate \( \text{CVR}_{\text{recom}}_{a,d,1} = \text{CVR}_{a,d} \). The outside option has a known conversion rate \( \text{CVR}_{\text{recom}}_{a,d,0} = \text{CVR}_{a,d} + s\Delta \), where at each replication we set \( s \in \{-1, 1\} \) with equal probabilities. The baseline conversion rate \( \text{CVR}_{\text{recom}}_{a,d} \) was calculated from the data, but is unknown to the recommender system. The parameter \( \Delta \) was set as a fraction of the average value of \( \text{CVR}_{\text{recom}}_{a,d} \) over our data set. The presented results obtained for \( \Delta \) being a fraction of 20% out of the average \( \text{CVR}_{\text{recom}}_{a,d} \), but we note that results are robust over the region \([5\%, 30\%]\) which we have experimented with. We assume that click-through rates are already known; for simplicity we set \( \text{CTR}_{a,d,1} = \text{CTR}_{a,d,0} = 0.05 \), but note that results are robust with respect to these quantities.

While the observations above are drawn randomly, the matrix \( H \) and sample path of auxiliary observations are fixed and determined from the data. We extracted the trajectory of information arrivals \( \{h_{1,t}\} \) from the number of readers that arrived to the article from an external search engine between consecutive decision epochs. Information flows in the data often include two or more auxiliary observations between consecutive decision periods; as we comment in §2.2, the performance bounds we establish in this paper indeed hold for any integer values assigned to entries of the matrix \( H \). For each epoch \( t \) and arrival-from-search \( m \in \{1, \ldots, h_{1,t}\} \) we denote by \( Y_{1,t,m} \in \{0,1\} \) an indicator of whether the reader clicked on a recommendation hosted by the article. We denote by \( \text{CVR}_{\text{search}}_{a,d} \) the fraction of readers that clicked on a recommendation hosted by article \( a \) during day \( d \) after arriving to that article from a search engine, and define:

\[
\alpha_{a,d} := \frac{\text{CVR}_{\text{recom}}_{a,d}}{\text{CVR}_{\text{search}}_{a,d}}.
\]

We construct an estimator of \( \alpha_{a,d} \) based on \( \alpha_{a,d-1} \), the fraction of the two conversion rates from the previous day. Note that, as a fraction, \( \alpha_{a,d-1} \) is not an unbiased estimator of \( \alpha_{a,d} \). We assume that \( (\alpha_{a,d-1}/\alpha_{a,d}) \) is a log-normal random variable, that is, \( \alpha_{a,d-1} = \alpha_{a,d} \cdot \exp \{ \tilde{\sigma}^2 W \} \) for \( W \sim \mathcal{N}(0,1) \) and \( \tilde{\sigma} > 0 \) that we estimated below, and construct the following unbiased estimator of \( \alpha_{a,d} \):

\[
\hat{\alpha}_{a,d} = \alpha_{a,d-1} \cdot \exp \left\{ -\frac{\tilde{\sigma}^2}{2} \right\}.
\]

We compared the Thompson Sampling variant given in §4.2 to a standard Thompson Sampling that ignores auxiliary observations. For a standard Thompson sampling, the variables in (2) are set as follows:

\[
n_{k,t} = \sum_{s=1}^{t-1} X_{1,t}^{(1)} \mathbb{1}\{\pi_s = k\}, \quad \bar{X}_{k,n_{k,t}} = \frac{\sum_{s=1}^{t-1} X_{1,s}^{(1)} \mathbb{1}\{\pi_s = k\} X_{1,t}^{(2)}}{\sum_{s=1}^{t-1} X_{1,t}^{(1)} \mathbb{1}\{\pi_s = k\}}
\]
and for the version of Thompson sampling that utilizes auxiliary observations, they are set as follows:

\[
\begin{align*}
    n_{k,t} := & \sum_{s=1}^{t-1} X_{1,t}^{(1)} 1\{\pi_s = k\} + \sum_{s=1}^{t} \frac{\sigma^2}{\hat{\sigma}^2} h_{1,s}, \\
\end{align*}
\]

\[
\begin{align*}
    \hat{X}_{k,n_{k,t}} := & \sum_{s=1}^{t} \frac{1}{\sigma^2} X_{1,t}^{(1)} 1\{\pi_s = k\} X_{k,s}^{(2)} + \sum_{s=1}^{t} \sum_{m=1}^{h_{1,t}} \frac{1}{\hat{\sigma}^2} Z_{1,t,m} \\
    \sum_{s=1}^{t-1} & \frac{1}{\sigma^2} X_{1,t}^{(1)} 1\{\pi_s = k\} + \sum_{s=1}^{t} \frac{1}{\hat{\sigma}^2} h_{1,s}
\end{align*}
\]

where \(\sigma^2 = 1/4\), \(\hat{\sigma}^2 = \alpha_{a,d-1}^2/4\), and \(Z_{1,t,m} = \hat{\alpha}_{a,d} Y_{k,t,m}\). Both algorithmic variants make decisions based on the rule:

\[
\pi_t = \begin{cases} 
0 & \text{if } \text{CVR}_{a,d,0}^{\text{recom}} > \theta_{1,t} \\
1 & \text{otherwise}.
\end{cases}
\]

Both algorithms where tuned using the parametric value \(c = 0.25\), which was identified through exhaustive search as the value that optimizes the performance of the standard Thompson sampling over the data set. We note that the performance of the Thompson sampling variant that utilizes information flows could be further improved by selecting a slightly lower value for this tuning parameter.

References

Adomavicius, G. and A. Tuzhilin (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering* (6), 734–749.

Agarwal, A., D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire (2014). Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646.

Agarwal, D., B.-C. Chen, P. Elango, N. Motgi, S.-T. Park, R. Ramakrishnan, S. Roy, and J. Zachariah (2009). Online models for content optimization. In *Advances in Neural Information Processing Systems*, pp. 17–24.

Agrawal, S. and N. Goyal (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pp. 1–39.

Agrawal, S. and N. Goyal (2013a). Further optimal regret bounds for thompson sampling. In *Artificial Intelligence and Statistics*, pp. 99–107.

Agrawal, S. and N. Goyal (2013b). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135.

Anderer, A., H. Bastani, and J. Silberholz (2019). Adaptive clinical trial designs with surrogates: When should we bother? *Available at SSRN*.

Ansari, A., S. Essegaier, and R. Kohli (2000). Internet recommendation systems. *Journal of Marketing Research* 37, 363–375.

Araman, V. F. and R. Caldentey (2009). Dynamic pricing for nonperishable products with demand learning. *Operations research* 57(5), 1169–1188.

Audibert, J.-Y. and S. Bubeck (2010). Best arm identification in multi-armed bandits. In *23th Conference on Learning Theory*.

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

73
Auer, P., N. Cesa-Bianchi, Y. Freund, and R. E. Schapire (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on, pp. 322–331.

Basilico, J. (2013). Recommendation at netflix scale. Slides, available at SlideShare: https://www.slideshare.net/justinbasilico/recommendation-at-netflix-scale.

Bastani, H. and M. Bayati (2015). Online decision-making with high-dimensional covariates. Preprint, available at SSRN: http://ssrn.com/abstract=2661896.

Bastani, H., M. Bayati, and K. Khosravi (2017). Exploiting the natural exploration in contextual bandits. arXiv preprint arXiv:1704.09011.

Bastani, H., P. Harsha, G. Perakis, and D. Singhvi (2018). Sequential learning of product recommendations with customer disengagement. Available at SSRN 3240970.

Bastani, H., D. Simchi-Levi, and R. Zhu (2019). Meta dynamic pricing: Learning across experiments. Available at SSRN 3334629.

Bergemann, D. and J. Välimäki (1996). Learning and strategic pricing. Econometrica: Journal of the Econometric Society, 1125–1149.

Berry, D. A. and B. Fristedt (1985). Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability), Volume 12. Springer.

Bertsimas, D., A. O’Hair, S. Relyea, and J. Silverholz (2016). An analytics approach to designing combination chemotherapy regimens for cancer. Management Science 62(5), 1511–1531.

Besbes, O., Y. Gur, and A. Zeevi (2016). Optimization in online content recommendation services: Beyond click-through rates. Manufacturing & Service Operations Management 18(1), 15–33.

Besbes, O., Y. Gur, and A. Zeevi (2018). Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. Forthcoming in Stochastic Systems.

Besbes, O. and A. Muharremoglu (2013). On implications of demand censoring in the newsvendor problem. Management Science 59(6), 1407–1424.

Boyd, S. and L. Vandenberghe (2004). Convex optimization. Cambridge university press.

Breese, J. S., D. Heckerman, and C. Kadie (1998). Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, pp. 43–52. Morgan Kaufmann Publishers Inc.

Bubeck, S. and N. Cesa-Bianchi (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends® in Machine Learning 5(1), 1–122.

Bubeck, S., V. Perchet, and P. Rigollet (2013). Bounded regret in stochastic multi-armed bandits. In Conference on Learning Theory, pp. 122–134.

Caro, F. and J. Gallien (2007). Dynamic assortment with demand learning for seasonal consumer goods. Management Science 53(2), 276–292.

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.

Cesa-Bianchi, N. and G. Lugosi (2006). Prediction, learning, and games. Cambridge university press.

Cesa-Bianchi, N., G. Lugosi, and G. Stoltz (2006). Regret minimization under partial monitoring. Mathematics of Operations Research 31(3), 562–580.
Chapelle, O. (2014). Modeling delayed feedback in display advertising. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1097–1105.

Dave, K. S. and V. Varma (2010). Learning the click-through rate for rare/new ads from similar ads. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 897–898.

den Boer, A. V. and B. Zwart (2013). Simultaneously learning and optimizing using controlled variance pricing. Management science 60(3), 770–783.

Farias, V. F. and A. A. Li (2019). Learning preferences with side information. Forthcoming in Management Science.

Farias, V. F. and B. Van Roy (2010). Dynamic pricing with a prior on market response. Operations Research 58(1), 16–29.

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 55(1), 119–139.

Gerchinovitz, S. and T. Lattimore (2016). Refined lower bounds for adversarial bandits. In Advances in Neural Information Processing Systems, pp. 1198–1206.

Gittins, J., K. Glazebrook, and R. Weber (2011). Multi-armed bandit allocation indices. John Wiley & Sons.

Goldenshluger, A. and A. Zeevi (2009). Woodroofe’s one-armed bandit problem revisited. The Annals of Applied Probability 19(4), 1603–1633.

Goldenshluger, A. and A. Zeevi (2013). A linear response bandit problem. Stochastic Systems 3(1), 230–261.

Gur, Y. and A. Momeni (2018). Adaptive learning with unknown information flows. Advances in Neural Information Processing Systems 31, 7473–7482.

Harrison, J. M., N. B. Keskin, and A. Zeevi (2012). Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. Management Science 58(3), 570–586.

Hill, W., L. Stead, M. Rosenstein, and G. Furnas (1995). Recommending and evaluating choices in a virtual community of use. In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 194–201. ACM Press/Addison-Wesley Publishing Co.

Huh, W. T. and P. Rusmevichientong (2009). A nonparametric asymptotic analysis of inventory planning with censored demand. Mathematics of Operations Research 34(1), 103–123.

Kleinberg, R. and T. Leighton (2003). The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on, pp. 594–605. IEEE.

Kohli, P., M. Salek, and G. Stoddard (2013). A fast bandit algorithm for recommendation to users with heterogeneous tastes. In Twenty-Seventh AAAI Conference on Artificial Intelligence.

Konstan, J. A., B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl (1997). GroupLens: applying collaborative filtering to usenet news. Communications of the ACM 40(3), 77–87.

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

Langford, J. and T. Zhang (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In Advances in neural information processing systems, pp. 817–824.

Lee, J., J. Lee, and H. Lee (2003). Exploration and exploitation in the presence of network externalities. Management Science 49(4), 553–570.
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.

Lee, K.-c., B. Orten, A. Dasdan, and W. Li (2012). Estimating conversion rate in display advertising from past performance data. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 768–776.

Li, L., W. Chu, J. Langford, and R. E. Schapire (2010). A contextual-bandit approach to personalized news article recommendation. In Proceedings of the 19th international conference on World Wide Web, pp. 661–670.

Madani, O. and D. DeCoste (2005). Contextual recommender problems. In Proceedings of the 1st international workshop on Utility-based data mining, pp. 86–89.

Nie, X., X. Tian, J. Taylor, and J. Zou (2017). Why adaptively collected data have negative bias and how to correct for it. arXiv preprint arXiv:1708.01977.

Pan, S. J. and Q. Yang (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering 22(10), 1345–1359.

Pandey, S., D. Agarwal, D. Chakrabarti, and V. Josifovski (2007). Bandits for taxonomies: A model-based approach. In Proceedings of the 2007 SIAM International Conference on Data Mining, pp. 216–227.

Park, S.-T. and W. Chu (2009). Pairwise preference regression for cold-start recommendation. In Proceedings of the third ACM conference on Recommender systems, pp. 21–28. ACM.

Regelson, M. and D. Fain (2006). Predicting click-through rate using keyword clusters. In Proceedings of the Second Workshop on Sponsored Search Auctions, Volume 9623, pp. 1–6.

Ricci, F., L. Rokach, and B. Shapira (2011). Introduction to recommender systems handbook. In Recommender systems handbook, pp. 1–35. Springer.

Richardson, M., E. Dominowska, and R. Ragno (2007). Predicting clicks: estimating the click-through rate for new ads. In Proceedings of the 16th international conference on World Wide Web, pp. 521–530.

Robbins, H. (1952). Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society 58(5), 527–535.

Rosales, R., H. Cheng, and E. Manavoglu (2012). Post-click conversion modeling and analysis for non-guaranteed delivery display advertising. In Proceedings of the fifth ACM international conference on Web search and data mining, pp. 293–302.

Sarkar, J. (1991). One-armed bandit problems with covariates. The Annals of Statistics 19(4), 1978–2002.

Schein, A. I., A. Popescul, L. H. Ungar, and D. M. Pennock (2002). Methods and metrics for cold-start recommendations. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 253–260. ACM.

Shah, V., J. Blanchet, and R. Johari (2018). Bandit learning with positive externalities. arXiv preprint arXiv:1802.05693.

Sharma, A., J. M. Hofman, and D. J. Watts (2015). Estimating the causal impact of recommendation systems from observational data. In Proceedings of the Sixteenth ACM Conference on Economics and Computation, pp. 453–470.

Strehl, A. L., C. Mesterharm, M. L. Littman, and H. Hirsh (2006). Experience-efficient learning in associative bandit problems. In Proceedings of the 23rd international conference on Machine learning, pp. 889–896.

Tang, L., Y. Jiang, L. Li, and T. Li (2014). Ensemble contextual bandits for personalized recommendation. In Proceedings of the 8th ACM Conference on Recommender Systems, pp. 73–80.
Tang, L., Y. Jiang, L. Li, C. Zeng, and T. Li (2015). Personalized recommendation via parameter-free contextual bandits. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, pp. 323–332.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika 25(3/4), 285–294.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. The Annals of Statistics 32(1), 135–166.

Tsybakov, A. B. (2009). Introduction to Nonparametric Estimation. Springer Publishing Company, Incorporated.

Villar, S. S., J. Bowden, and J. Wason (2015). Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. Statistical science: a review journal of the Institute of Mathematical Statistics 30(2), 199.

Wang, C.-C., S. R. Kulkarni, and H. V. Poor (2005). Bandit problems with side observations. IEEE Transactions on Automatic Control 50(3), 338–355.

Wang, L., C. Wang, K. Wang, and X. He (2017). Biucb: A contextual bandit algorithm for cold-start and diversified recommendation. In 2017 IEEE International Conference on Big Knowledge (ICBK), pp. 248–253.

Woodroofe, M. (1979). A one-armed bandit problem with a concomitant variable. Journal of the American Statistical Association 74(368), 799–806.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. Journal of the American Statistical Association 64(325), 131–146.

Zhang, J., P. W. Farris, J. W. Irvin, T. Kushwaha, T. J. Steenburgh, and B. A. Weitz (2010). Crafting integrated multichannel retailing strategies. Journal of Interactive Marketing 24(2), 168–180.

Zhang, W., S. Yuan, J. Wang, and X. Shen (2014). Real-time bidding benchmarking with ipinyou dataset. arXiv preprint arXiv:1407.7073.