Asymptotically and Minimax Optimal Regret Bounds for Multi-Armed Bandits with Abstention

Junwen Yang *junwen_yang@u.nus.edu*
Institute of Operations Research and Analytics
National University of Singapore

Tianyuan Jin *tianyuan@u.nus.edu*
Department of Electrical and Computer Engineering
National University of Singapore

Vincent Y. F. Tan *vtan@nus.edu.sg*
Department of Mathematics
Department of Electrical and Computer Engineering
Institute of Operations Research and Analytics
National University of Singapore

Reviewed on OpenReview: *https://openreview.net/forum?id=AYp5zOcFdA*

Abstract

We introduce a novel extension of the canonical multi-armed bandit problem that incorporates an additional strategic innovation: *abstention*. In this enhanced framework, the agent is not only tasked with selecting an arm at each time step, but also has the option to *abstain* from accepting the stochastic instantaneous reward before observing it. When opting for abstention, the agent either suffers a fixed regret or gains a guaranteed reward. This added layer of complexity naturally prompts the key question: can we develop algorithms that are both computationally efficient and asymptotically and minimax optimal in this setting? We answer this question in the affirmative by designing and analyzing algorithms whose regrets meet their corresponding information-theoretic lower bounds. Our results offer valuable quantitative insights into the benefits of the abstention option, laying the groundwork for further exploration in other online decision-making problems with such an option. Extensive numerical experiments validate our theoretical results, demonstrating that our approach not only advances theory but also has the potential to deliver significant practical benefits.

1 Introduction

In online decision-making, the multi-armed bandit model, originally introduced by Thompson (1933), has long served as a quintessential benchmark for capturing the delicate interplay between exploration and exploitation. In stochastic multi-armed bandit problems, the agent sequentially selects an arm from the given set at each time step and subsequently observes a random reward associated with the chosen arm. To maximize cumulative rewards, the agent must strike a balance between the persistent pursuit of the arm with the highest estimated reward (exploitation) and the adventurous exploration of other arms to gain a deeper understanding of their potential (exploration). This fundamental challenge finds applications across a wide array of domains, ranging from optimizing advertising campaigns to fine-tuning recommendation systems.

However, real-world scenarios often come fraught with complexities that challenge the simplicity of the canonical bandit model. One notable complexity arises when the agent is equipped with an additional option to abstain from accepting the stochastic instantaneous reward before observing it. This added layer of decision-making enriches the strategic landscape, altering how the agent optimally navigates the trade-off between exploration and exploitation.

Consider clinical trials, where concerns about participant harm, insurance, and compensation are prominent issues discussed in the literature; see, for example, Steinbrook (2006); Munshi & Thatte (2013).[1] No-fault compensation or prophylactic measures (Henry et al., 2015) allow researchers to observe the response to a potentially hazardous treatment while avoiding exposure to random harmful effects. When evaluating potentially hazardous medical treatments, researchers can proactively deploy safeguards such as preemptive medications or rely on arrangements for insurance or compensation against possible research-related injuries. However, these protective measures come with costs, which may be modeled as either fixed regrets or fixed rewards in the context of the clinical study's cumulative regret. In these scenarios, researchers have the option to observe the outcomes of a treatment while abstaining from incurring the associated random regret through these costly prearranged measures. Subsequently, these observed outcomes can be leveraged to inform more efficacious future experimental designs. Consequently, opting for abstention has the potential to promote more responsible decision-making and reduce the overall cumulative regret of the study.

To further illustrate our motivation, consider another example in online advertising, where companies frequently grapple with the challenge of allocating their advertising budget optimally across multiple platforms (Google Ads, LinkedIn, etc.) to promote their products. This scenario naturally parallels the framework of multi-armed bandits, with each arm representing a distinct advertising platform. Typically, these platforms operate on a *cost-per-click* pricing model, while performance-based alternatives such as *cost-per-action* (CPA) pricing have been extensively studied as a mechanism to shift the risk of uncertain consumer behavior from the advertiser to the platform (Nazerzadeh et al., 2008; Hu et al., 2016). As detailed by Hu et al. (2016), CPA models act as a financial safeguard for advertisers by guaranteeing that payments are only exacted upon successful downstream conversions, rather than merely clicks. However, what companies truly care about is the outcome, such as conversions or purchases, which are random variables sampled from unknown distributions. When faced with such uncertainty and the possibility of underperformance, companies may instead opt for a *cost-per-action* strategy, which involves a predetermined payment for a specific outcome (i.e., payment is made only when a desired outcome, such as a conversion or purchase, occurs). This strategy essentially serves as an abstention option, where instead of risking uncertain returns (from unknown distributions), companies choose to pay a fixed cost for a known outcome. By embracing the abstention option, companies can mitigate the uncertainties associated with stochastic outcomes and ensure a deterministic result. Moreover, by opting for abstention, the company still retains the ability to observe the advertising conversion rate of the platform, enabling them to gather valuable insights for future decision-making.

Building upon this challenge, we introduce an innovative extension to the canonical multi-armed bandit model that incorporates abstention as a legitimate strategic option. At each time step, the agent not only selects which arm to pull but also decides whether to abstain. Depending on how the abstention option impacts the cumulative regret, which is the agent's primary optimization objective, our abstention model offers two complementary settings, namely, the fixed-regret setting where abstention results in a constant regret, and the fixed-reward setting where abstention yields a deterministic reward. Collectively, these settings provide the agent with a comprehensive toolkit for adeptly navigating the complicated landscape of online decision-making.

Main contributions. Our main results and contributions are summarized as follows:

(i) In Section 2, we provide a rigorous formulation of the multi-armed bandit model with abstention. Our focus is on cumulative regret minimization across two distinct yet complementary settings: *fixed-regret* and *fixed-reward*. These settings give rise to different metrics, each offering unique analytical insights. Importantly, both settings encompass the canonical bandit model as a particular case.

(ii) In the fixed-regret setting, we judiciously combine two abstention criteria into a Thompson Sampling-based algorithm proposed by Jin et al. (2023). This integration ensures compatibility with the abstention option, as elaborated in Algorithm 1. The first abstention criterion employs a carefully constructed lower confidence bound, while the second is tailored to mitigate worst-case scenarios.

[1] We highlight that this example is a stylized mathematical abstraction used solely to motivate the abstention framework. It does not capture the complex ethical and practical constraints of actual clinical trials and should not be interpreted as a proposal for real-world medical protocols.

We establish both asymptotic and minimax upper bounds on the cumulative regret, requiring application of analytical techniques inspired by both Thompson Sampling and UCB methodologies. Furthermore, we derive corresponding lower bounds, thereby showcasing the concurrent attainment of asymptotic and minimax optimality by our algorithm. Our approach entails extensive application of analytical techniques inspired by both Thompson Sampling and UCB methodologies.

(iii) In the fixed-reward setting, we introduce a general strategy, outlined in Algorithm 2. This method transforms any algorithm that is both asymptotically and minimax optimal in the canonical model to one that also accommodates the abstention option. Remarkably, this strategy maintains its universal applicability and straightforward implementation while provably achieving both forms of optimality—asymptotic and minimax.

(iv) To corroborate our theoretical contributions, we conduct a series of numerical experiments in Appendix F. These experiments substantiate the effectiveness of our algorithms and highlight the performance gains achieved through the inclusion of the abstention option.

1.1 Related Work

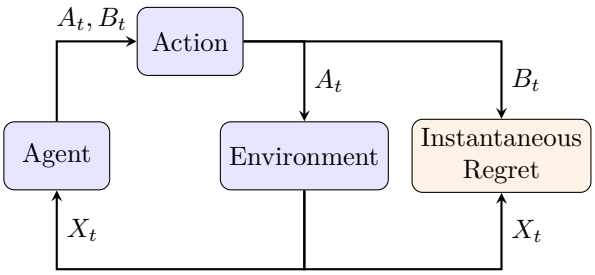
Canonical multi-armed bandits. The study of cumulative regret minimization in canonical multi-armed bandits has attracted considerable scholarly focus. Within this domain, two dominant paradigms for evaluating optimality metrics have emerged: asymptotic optimality and minimax optimality. In essence, the former considers the behavior of algorithms as the time horizon approaches infinity for a specific problem instance, while the latter seeks to minimize the worst-case regret over all possible instances. A diverse array of policies have been rigorously established to achieve asymptotic optimality across various model settings. Notable examples include UCB2 (Auer et al., 2002), DMED (Honda & Takemura, 2010), KL-UCB (Cappé et al., 2013), and Thompson Sampling (Agrawal & Goyal, 2012; Kaufmann et al., 2012). In the context of the worst-case regret, MOSS (Audibert & Bubeck, 2009) stands out as the pioneering method that has been verified to be minimax optimal. Remarkably, KL-UCB^{++} (Ménard & Garivier, 2017) became the first algorithm proved to achieve both asymptotic and minimax optimality. Recently, Jin et al. (2023) introduced Less-Exploring Thompson Sampling, an innovation that boosts computational efficiency compared to classical Thompson Sampling while concurrently achieving asymptotic and minimax optimality. For a comprehensive survey of bandit algorithms, we refer to Lattimore & Szepesvári (2020).

Machine learning with abstention. Starting with the seminal work of Chow (1957; 1970), the concept of learning with *abstention* (also referred to as *rejection*) has been extensively explored in various machine learning paradigms. These include classification (Herbei & Wegkamp, 2006; Bartlett & Wegkamp, 2008; Cortes et al., 2016), ranking (Cheng et al., 2010; Mao et al., 2023) and regression (Wiener & El-Yaniv, 2012; Zaoui et al., 2020; Kalai & Kanade, 2021).

Within this broad spectrum of research, our work is most directly related to those that explore the role of abstention in the context of *online learning*. To the best of our knowledge, Cortes et al. (2018) firstly incorporated the abstention option into the problem of online prediction with expert advice (Littlestone & Warmuth, 1994). In their model, at each time step, each expert has the option to either make a prediction based on the given input or abstain from doing so. When the agent follows the advice of an expert who chooses to abstain, the true label of the input remains undisclosed, and the learner incurs a known fixed loss. Subsequently, Neu & Zhivotovskiy (2020) introduced a different abstention model, which is more similar to ours. Here, the abstention option is only available to the agent. Crucially, the true label is always revealed to the agent *after* the decision has been made, regardless of whether the agent opts to abstain. Their findings suggest that equipping the agent with an abstention option can significantly improve the guarantees on the worst-case regret.

2 Problem Setup

Multi-armed bandits with abstention. We consider a K-armed bandit model, enhanced with an additional option to abstain from accepting the stochastic instantaneous reward prior to its observation. Let



	$B_t = 0$	$B_t = 1$
Fixed-Regret	$\mu_1 - X_t$	c
Fixed-Reward	$\mu_1 \vee c - X_t$	$\mu_1 \vee c - c$

Figure 1: Interaction protocol for multi-armed bandits with fixed-regret and fixed-reward abstention.

$\mu \in \mathcal{U} := \mathbb{R}^K$ denote a specific bandit instance, where μ_i represents the unknown mean reward of arm $i \in [K]$. We assume that arm 1 is the unique optimal arm, i.e., $1 = \arg\max_{i\in[K]} \mu_i$, and we define $\Delta_i := \mu_1 - \mu_i$ as the suboptimality gap for each arm i.

At each time step $t \in \mathbb{N}$, the agent chooses an arm $A_t \in [K]$, and, simultaneously, decides whether or not to abstain, indicated by a binary variable B_t. Regardless of the decision to abstain, the agent observes a random variable X_t from the selected arm A_t, which is drawn from a Gaussian distribution $\mathcal{N}(\mu_{A_t}, 1)$ independent of observations obtained from the previous time steps. Notably, the selection of both A_t and B_t might depend on the previous decisions and observations, as well as on each other. Formally, let $\mathcal{F}_t := \sigma(A_1, B_1, X_1, \ldots, A_t, B_t, X_t)$ denote the σ-field generated by the cumulative interaction history up to and including time t. It follows that the pair of random variables (A_t, B_t) is $\mathcal{F}_{t-1}$-measurable.

The instantaneous regret at time t is determined by both the binary abstention variable B_t and the observation X_t. Based on the outcome of the abstention option, we now discuss two complementary settings. In the *fixed-regret* setting, the abstention option incurs a constant regret. Opting for abstention ($B_t = 1$) leads to a deterministic regret of $c > 0$, in contrast to the initial regret linked to arm A_t when not selecting abstention ($B_t = 0$), which is given by $\mu_1 - X_t$.

In the *fixed-reward* setting, the reward of the abstention option is $c \in \mathbb{R}$.[2] Since the abstention reward c may be larger than μ_1, the highest expected reward at each time step is $\mu_1 \vee c := \max\{\mu_1, c\}$. If the agent decides to abstain ($B_t = 1$), it is guaranteed a deterministic reward of c, leading to a regret of $\mu_1 \vee c - c$. If $B_t = 0$, the agent receives a reward of X_t, resulting in a regret of $\mu_1 \vee c - X_t$.

See Figure 1 for a schematic of our model in the two settings.

Remark 1. We note that if the agent cannot observe the sample from the chosen arm when opting for abstention, then it is equivalent to skipping the time step, rendering the model trivial. In other words, the learner gains no information from the time step but incurs an instantaneous regret. In view of this, our model well aligns with that of Neu & Zhivotovskiy (2020), which explored the role of abstention in the context of online prediction with expert advice.

Remark 2 (Graph Feedback). Bandits with graph feedback (Liu et al., 2018; Alon et al., 2017) are a natural extension of the standard multi-armed bandit setting. In this framework, each arm is represented as a node in a graph, and pulling one arm reveals not only its own reward but also observations from the neighboring nodes connected to it in the graph. Our model bears a conceptual resemblance to bandits with graph feedback: it can be represented as a bipartite graph in which each stochastic arm a is associated with a corresponding abstention arm Q_a. Pulling a stochastic arm a reveals a reward drawn from arm a's distribution, while choosing the corresponding abstention arm Q_a yields an observation drawn from the reward distribution of the associated stochastic arm as well.

That said, there is a key structural difference between the two frameworks. In bandits with graph feedback, selecting an arm reveals not only the reward of the chosen arm but also observations from its neighboring arms. In contrast, in our model, choosing an abstention arm does not provide any information about the

[2]With a slight abuse of notation, we use the symbol c to represent the abstention regret and the abstention reward within their respective settings. The surrounding context should elucidate the exact meaning of c.

abstention arm itself. Instead, the mean reward (or regret) of each abstention arm is known to the learner at the beginning of the game.

Regret minimization. Our overarching goal is to design and analyze online algorithms π that minimize their expected cumulative regrets up to and including the time horizon T.[3] The regrets are formally defined for the two distinct settings as follows:

- Fixed-regret setting:

$$R^{\mathrm{RG}}_{\mu,c}(T,\pi) := \mathbb{E}\left[\sum_{t=1}^{T}\big((\mu_1 - X_t)\cdot \mathbf{1}\{B_t=0\} + c\cdot \mathbf{1}\{B_t=1\}\big)\right]. \tag{1}$$

- Fixed-reward setting:

$$R^{\mathrm{RW}}_{\mu,c}(T,\pi) := T\cdot(\mu_1 \vee c) - \mathbb{E}\left[\sum_{t=1}^{T}\Big(X_t\cdot \mathbf{1}\{B_t=0\} + c\cdot \mathbf{1}\{B_t=1\}\Big)\right]. \tag{2}$$

An online algorithm π consists of two interrelated components: the *arm sampling* rule that selects A_t, and the *abstention decision rule* that determines B_t at each time step $t \in [T]$. Additionally, we use Π^{RG} and Π^{RW} to denote the collections of all online policies for the fixed-regret and fixed-reward settings, respectively. For the sake of analytical convenience, we also introduce the canonical regret $R^{\mathrm{CA}}_{\mu}(T,\pi) := T\mu_1 - \mathbb{E}\big[\sum_{t=1}^{T} X_t\big]$, which disregards the abstention option and remains well-defined within our abstention model. Furthermore, when there is no ambiguity, we will omit the dependence of the regret on the policy. For example, we often abbreviate $R^{\mathrm{RG}}_{\mu,c}(T,\pi)$ as $R^{\mathrm{RG}}_{\mu,c}(T)$.

Remark 3. It is worth mentioning that our model is a strict generalization of the canonical multi-armed bandit model (without the abstention option). Specifically, it particularizes to the canonical model as the abstention regret c tends to $+\infty$ in the fixed-regret setting and as the abstention reward c tends to $-\infty$ in the fixed-reward setting. Nevertheless, the incorporation of an extra challenge, the abstention decision (denoted as B_t), offers the agent the potential opportunity to achieve superior performance in terms of either regret. One might wonder whether our setting reduces to a $(K+1)$-armed bandit model where the $(K+1)$-st arm models abstention. It does not; for a detailed discussion on why our framework cannot be reduced to a standard bandit problem with an additional deterministic arm, the reader is referred to Appendix B.

Other notations. For $x, y \in \mathbb{R}$, we denote $x \wedge y := \min\{x,y\}$ and $x \vee y := \max\{x,y\}$. For any arm $i \in [K]$, let $N_i(t) := \sum_{s=1}^{t}\mathbf{1}\{A_s = i\}$ and $\hat{\mu}_i(t) := \sum_{s=1}^{t} X_s\mathbf{1}\{A_s = i\}/N_i(t)$ denote its total number of pulls and empirical estimate of the mean up to time t, respectively. In particular, we set $\hat{\mu}_i(t) = +\infty$ if $N_i(t) = 0$. To count abstention records, we also use $N_i^{(0)}(t)$ and $N_i^{(1)}(t)$ to denote its number of pulls without and with abstention up to time t, respectively. That is, $N_i^{(0)}(t) := \sum_{s=1}^{t}\mathbf{1}\{A_s = i \text{ and } B_s = 0\}$ and $N_i^{(1)}(t) := \sum_{s=1}^{t}\mathbf{1}\{A_s = i \text{ and } B_s = 1\}$. Additionally, we define $\hat{\mu}_{is}$ as the empirical mean of arm i based on its first s pulls. Furthermore, we use α, α_1, and so forth to represent universal constants that do not depend on the *problem instances* (including μ, c, T, K), with possibly different values in different contexts.

3 Fixed-Regret Setting

In this section, we focus on the fixed-regret setting. Specifically, we design a conceptually simple and computationally efficient algorithm, namely Fixed-Regret Thompson Sampling with Abstention (or FRG-TSwA), to minimize the cumulative regret while incorporating fixed-regret abstention. To evaluate the performance of our algorithm from a theoretical standpoint, we establish both instance-dependent asymptotic and instance-independent minimax upper bounds on the cumulative regret, as elaborated upon in Section 3.1. Furthermore, in Section 3.2, we provide lower bounds for the problem of regret minimization in multi-armed bandits with fixed-regret abstention. These findings substantiate that our algorithm achieves both asymptotic

[3] In certain real-world applications, the time horizon T may be unknown to the agent. In fact, all of our proposed methods are inherently *anytime* in nature, as they do not necessitate prior knowledge of the horizon.

Algorithm 1 Fixed-Regret Thompson Sampling with Abstention (or FRG-TSwA)

Input: Arm set $[K]$ and abstention regret $c > 0$.

1: Sample each arm once, and choose to abstain ($B_t = 1$) if and only if $\sqrt{\frac{K}{t}} \geq c$.
2: Initialize $\hat{\mu}_i(K)$ and $N_i(K) = 1$ for all $i \in [K]$.
3: **for** $t = K+1, \ldots, T$ **do**
4: For each arm $i \in [K]$, sample $\theta_i(t) \sim \mathcal{N}(\hat{\mu}_i(t-1), 1/N_i(t-1))$ and set

$$a_i(t) = \begin{cases} \theta_i(t) & \text{with probability } 1/K \\ \hat{\mu}_i(t-1) & \text{with probability } 1 - 1/K. \end{cases}$$

5: Pull the arm $A_t = \arg\max_{i \in [K]} a_i(t)$, and choose to abstain ($B_t = 1$) if and only if

$$\max_{i \in [K] \setminus \{A_t\}} \mathrm{LCB}_i(t) - \hat{\mu}_{A_t}(t-1) \geq c \quad \text{or} \quad \sqrt{\frac{K}{t}} \geq c.$$

6: Observe X_t from the arm A_t, and update $\hat{\mu}_i(t)$ and $N_i(t)$ for all $i \in [K]$.
7: **end for**

and minimax optimality simultaneously. The pseudocode for FRG-TSwA is presented in Algorithm 1 and elucidated in the following.

In terms of the arm sampling rule, our algorithm is built upon Less-Exploring Thompson Sampling (Jin et al., 2023), a minimax optimal enhancement of the celebrated Thompson Sampling (TS) algorithm (Thompson, 1933). We refer to Remark 4 for the reason behind this choice. During the initialization phase, each arm is sampled exactly once. Following that, at each time t, an estimated reward $a_i(t)$ is constructed for each arm $i \in [K]$, which is either drawn from the posterior distribution $\mathcal{N}(\hat{\mu}_i(t-1), 1/N_i(t-1))$ with probability $1/K$ or set to be the empirical mean $\hat{\mu}_i(t-1)$ otherwise. Subsequently, the algorithm consistently pulls the arm A_t with the highest estimated reward.

To pursue the dual optimality properties, it is essential to carefully balance the simultaneous control of asymptotic and worst-case regrets resulting from potential misjudgments in choosing the abstention option. With regard to the abstention decision rule, we propose two abstention criteria that work in tandem (as detailed in Step 5 of Algorithm 1). The first criterion is gap-dependent in nature. In particular, for each arm $i \in [K]$, we define its *lower confidence bound* as

$$\mathrm{LCB}_i(t) := \hat{\mu}_i(t-1) - \sqrt{\frac{6 \log t + 2 \log(c \vee 1)}{N_i(t-1)}}. \tag{3}$$

Then we choose to abstain if there exists an arm $i \in [K] \setminus \{A_t\}$ for which the difference between $\mathrm{LCB}_i(t)$ and the empirical mean of the arm A_t exceeds c. This condition signifies that the suboptimality gap Δ_{A_t} is at least c with high probability. The LCB in (3) is custom-tailored for our specific application, which is unique in the fact that it takes the value of the abstention regret c into consideration. Notably, besides its asymptotic optimality, our choice to employ this LCB rather than other natural types of upper or lower confidence bounds in the abstention criteria (e.g., those that do not include c) is primarily due to the imperative to constrain the minimax regret to be $O(\sqrt{KT})$.

The second abstention criterion is gap-independent. It is motivated from the construction of worst-case scenarios as detailed in the proof of our lower bound. Under this criterion, we opt for the abstention option if $c \leq \sqrt{K/t}$, which implies that the abstention regret remains acceptably low at time t in view of the worst-case scenarios.

Remark 4. As previously highlighted, our model in the fixed-regret setting particularizes to the canonical multi-armed bandit model as the abstention regret c tends to infinity. Similarly, when c tends to infinity, the two abstention criteria are never satisfied, and the procedure of Algorithm 1 simplifies to that of Less-

Exploring Thompson Sampling. It is worth noting that this latter algorithm is not only asymptotically optimal but also minimax optimal for the canonical model. This is precisely why we base our algorithm upon it, rather than the conventional Thompson Sampling algorithm, which has been shown not to be minimax optimal (Agrawal & Goyal, 2017).

3.1 Upper Bounds

Theorem 1 below provides two distinct types of theoretical guarantees pertaining to our algorithm's performance on the cumulative regret $R^{\mathrm{RG}}_{\mu,c}(T)$, defined in Equation (1) for the fixed-regret setting.

Theorem 1. *For all abstention regrets $c > 0$ and bandit instances $\mu \in \mathcal{U}$, Algorithm 1 guarantees that*

$$\limsup_{T\to\infty} \frac{R^{\mathrm{RG}}_{\mu,c}(T)}{\log T} \leq 2\sum_{i>1} \frac{\Delta_i \wedge c}{\Delta_i^2}.$$

Furthermore, there exists a universal constant $\alpha > 0$ such that

$$R^{\mathrm{RG}}_{\mu,c}(T) \leq \begin{cases} cT & \text{if } c \leq \sqrt{K/T} \\ \alpha(\sqrt{KT} + \sum_{i>1} \Delta_i) & \text{if } c > \sqrt{K/T}. \end{cases}$$

The proof of Theorem 1 is deferred to Appendix D.1. The theoretical challenges arising from proving Theorem 1 revolve around the quantification of the regret that results from inaccurately estimating the suboptimality gaps associated with the abstention criteria. More precisely, it is crucial to establish upper bounds on $\mathbb{E}[N_i^{(1)}(T)]$ for arms i with $\Delta_i < c$ (which, by definition, includes the best arm), and on $\mathbb{E}[N_i^{(0)}(T)]$ for arms i with $\Delta_i > c$. These bounds need to be derived from both asymptotic and minimax perspectives, adding layers of complexity to the analytical process. To accomplish these tasks, we initially decompose the target expectations into manageable components, and subsequently bound the resulting subterms in the two respective regimes. Our approach, merging TS-based sampling rule with LCB-based abstention criteria, necessitates a careful amalgamation of both TS-type and UCB-type analytical techniques.

Furthermore, for the asymptotic upper bound, we take a deeper exploration into the randomized arm sampling dynamics inherent to Less-Exploring Thompson Sampling. A pivotal aspect of this exploration is a characterization of a high probability lower bound on the number of pulls of the optimal arm, as detailed in Lemma 5, inspired by the work of Korda et al. (2013).

Finally, we remark that the numerous intricacies involved in these analyses preclude us from formulating a generalized strategy akin to the forthcoming Algorithm 2 for the fixed-reward setting.

3.2 Lower Bounds

In order to establish the asymptotic lower bound, we need to introduce the concept of R^{RG}-consistency, which rules out overly specialized algorithms that are tailored exclusively to specific problem instances. Roughly speaking, a R^{RG}-consistent algorithm guarantees a subpolynomial cumulative regret for any given problem instance.

Definition 1 (R^{RG}-consistency)**.** An algorithm $\pi \in \Pi^{\mathrm{RG}}$ is said to be R^{RG}*-consistent* if for all abstention regrets $c > 0$, bandit instances $\mu \in \mathcal{U}$, and $a > 0$, $R^{\mathrm{RG}}_{\mu,c}(T,\pi) = o(T^a)$.

Now we present both asymptotic and minimax lower bounds on the cumulative regret in Theorem 2, which is proved in Appendix D.2.

Theorem 2. *For any abstention regret $c > 0$, bandit instance $\mu \in \mathcal{U}$ and R^{RG}-consistent algorithm π, it holds that*

$$\liminf_{T\to\infty} \frac{R^{\mathrm{RG}}_{\mu,c}(T,\pi)}{\log T} \geq 2\sum_{i>1} \frac{\Delta_i \wedge c}{\Delta_i^2}.$$

For any abstention regret $c > 0$ and time horizon $T \geq K$, there exists a universal constant $\alpha > 0$ such that

$$\inf_{\pi\in\Pi^{\mathrm{RG}}} \sup_{\mu\in\mathcal{U}} R^{\mathrm{RG}}_{\mu,c}(T,\pi) \geq \alpha(\sqrt{KT} \wedge cT).$$

Comparing the upper bounds on the cumulative regret of our algorithm FRG-TSwA in Theorem 1 with the corresponding lower bounds in Theorem 2, it is evident that our algorithm exhibits both asymptotic and minimax optimality.

Asymptotic optimality. For any abstention regret $c > 0$ and bandit instance $\mu \in \mathcal{U}$, the regret of our algorithm satisfies the following limiting behaviour:

$$\lim_{T\to\infty} \frac{R^{\mathrm{RG}}_{\mu,c}(T)}{\log T} = 2\sum_{i>1} \frac{\Delta_i \wedge c}{\Delta_i^2}.$$

The above asymptotically optimal result yields several intriguing implications. First, the inclusion of the additional fixed-regret abstention option does not obviate the necessity of differentiating between suboptimal arms and the optimal one, and the exploration-exploitation trade-off remains crucial. In fact, to avoid the case in which the cumulative regret grows polynomially, the agent must still asymptotically allocate the same proportion of pulls to each suboptimal arm, as in the canonical model. This assertion is rigorously demonstrated in the proof of the lower bound (refer to Appendix D.2 for details). Nevertheless, the abstention option does indeed reduce the exploration cost for the agent. Specifically, when exploring any suboptimal arm with a suboptimality gap larger than c, our algorithm tends towards employing the abstention option to minimize the instantaneous regret. This aspect is formally established in the proof of the asymptotic upper bound (see Appendix D.1).

Minimax optimality. In the context of worst-case guarantees for the cumulative regret, we focus on the dependence on c, K, and T. Notably, the $\sum_{i>1} \Delta_i$ term[4] is typically considered as negligible in the literature (Audibert & Bubeck, 2009; Agrawal & Goyal, 2017; Lattimore & Szepesvári, 2020). Therefore, Theorem 1 demonstrates that our algorithm attains a worst-case regret of $O(\sqrt{KT} \wedge cT)$, which is minimax optimal in light of Theorem 2.

A phase transition phenomenon can be clearly observed from the worst-case guarantees, which dovetails with our intuitive understanding of the fixed-regret abstention setting. When the abstention regret c is sufficiently low, it becomes advantageous to consistently opt for abstention to avoid the worst-case scenarios. On the contrary, when the abstention regret c exceeds a certain threshold, the abstention option proves to be inadequate in alleviating the worst-case regret, as compared to the canonical model.

Remark 5. Although our model allows for the selected arm A_t and the abstention option B_t to depend on each other, the procedure used in both algorithms within this work is to first determine A_t *before* B_t; this successfully achieves both forms of optimality. Nevertheless, this approach might no longer be optimal beyond the canonical K-armed bandit setting. In K-armed bandits, each arm operates independently. Conversely, in models like linear bandits, pulling one arm can indirectly reveal information about other arms. Policies based on the principle of optimism in the face of uncertainty, as well as Thompson Sampling, fall short of achieving asymptotic optimality in the context of linear bandits (Lattimore & Szepesvári, 2017). Therefore, the abstention option becomes particularly attractive if there exists an arm that incurs a substantial regret but offers significant insights into the broader bandit instance.

Stochastic and heterogeneous abstention regret. If the regret incurred by the agent when opting for abstention is not deterministic but stochastically generated from a distribution with known expectation c, Algorithm 1 remains effective, and our analyses of the upper and lower bounds remain valid. Additionally, another natural extension of our model is that the abstention regret is different for different arms. Specifically, the agent incurs a regret of $c_i > 0$ when pulling arm i and choosing abstention. In this scenario, the abstention criteria in Algorithm 1 can be adapted accordingly, and the same asymptotic and minimax optimality can be established; see Appendix D.3 for details.

[4]This term is unavoidable when the abstention regret c is sufficiently high, since every reasonable algorithm has to allocate a fixed number of pulls to each arm.

Algorithm 2 Fixed-Reward Algorithm with Abstention (or FRW-ALGwA)

Input: Arm set $[K]$, abstention reward $c \in \mathbb{R}$, and a base algorithm ALG that is both asymptotically and minimax optimal for the canonical multi-armed bandit model.

Initialize $\hat{\mu}_i(0) = +\infty$ for all arms $i \in [K]$.
for $t = 1, 2, \ldots, T$ **do**
  Pull the arm A_t chosen by the base algorithm ALG.
  Choose to abstain ($B_t = 1$) if and only if $\hat{\mu}_{A_t}(t-1) \leq c$.
  Observe X_t from the arm A_t, and update $\hat{\mu}_i(t)$ for all $i \in [K]$.
end for

4 Fixed-Reward Setting

In this section, we investigate the fixed-reward setting. Here, the reward associated with the abstention option remains consistently fixed at $c \in \mathbb{R}$. When exploring a specific arm, the agent has the capability to determine whether selecting the abstention option yields a higher reward *solely* based on its own estimated mean reward. However, in the fixed-regret setting, this decision can only be made by taking into account *both* its own estimated mean reward and the estimated mean reward of the potentially best arm. In this regard, the fixed-reward setting is inherently less complex than the fixed-regret setting. As a result, it becomes possible for us to design a more general strategy Fixed-Reward Algorithm with Abstention (or FRW-ALGwA), whose pseudocode is presented in Algorithm 2. Despite the straightforward nature of our algorithm, we demonstrate its dual attainment of both asymptotic and minimax optimality through an exhaustive theoretical examination in Sections 4.1 and 4.2.

FRW-ALGwA leverages a base algorithm ALG that is asymptotically and minimax optimal for canonical multi-armed bandits as its input. For comprehensive definitions of asymptotic and minimax optimality within the canonical model, see Appendix A. Notably, eligible candidate algorithms include KL-UCB^{++} (Ménard & Garivier, 2017), ADA-UCB (Lattimore, 2018), MOTS-$\mathcal{J}$ (Jin et al., 2021) and Less-Exploring Thompson Sampling (Jin et al., 2023). In our algorithm, at each time step t, the base algorithm determines the selected arm A_t according to the partial interaction historical information $(A_1, X_1, A_2, X_2, \ldots, A_{t-1}, X_{t-1})$. Subsequently, the algorithm decides whether or not to abstain, indicated by the binary random variable B_t, by comparing the empirical mean of the arm A_t, denoted as $\hat{\mu}_{A_t}(t-1)$, to the abstention reward c.

4.1 Upper Bounds

Recall the definition of the cumulative regret $R^{\mathrm{RW}}_{\mu,c}(T)$, as presented in Equation (2) for the fixed-reward setting. Theorem 3 establishes both the instance-dependent asymptotic and instance-independent minimax upper bounds for Algorithm 2; see Appendix E.1 for the proof.

Theorem 3. *For all abstention rewards $c \in \mathbb{R}$ and bandit instances $\mu \in \mathcal{U}$, Algorithm 2 guarantees that*

$$\limsup_{T\to\infty} \frac{R^{\mathrm{RW}}_{\mu,c}(T)}{\log T} \leq 2 \sum_{i>1} \frac{\mu_1 \vee c - \mu_i \vee c}{\Delta_i^2}.$$

Furthermore, there exists a universal constant $\alpha > 0$ such that

$$R^{\mathrm{RW}}_{\mu,c}(T) \leq \alpha \left(\sqrt{KT} + \sum_{i\in[K]} (\mu_1 \vee c - \mu_i) \right).$$

Remark 6. It is worth considering the special case where $c \geq \mu_1$, where opting for abstention results in a reward even greater than, or equal to, the mean reward of the best arm. For this particular case, as per Theorem 3, since $\mu_1 \vee c - \mu_i \vee c = 0$ for all $i > 1$, our algorithm achieves a regret of $o(\log T)$. This result, in fact, is not surprising. In contrast to the fixed-regret setting where the regret associated with the abstention option is strictly positive, in this specific scenario of the fixed-reward setting, selecting the abstention option is indeed the optimal action at a single time step, regardless of the arm pulled. Therefore, there is no

necessity to distinguish between suboptimal arms and the optimal one, and the exploration-exploitation trade-off becomes inconsequential. However, when the abstention reward is below the mean reward of the best arm, i.e., $c < \mu_1$, maintaining a subpolynomial cumulative regret still hinges on the delicate balance between exploration and exploitation, as evidenced by the forthcoming exposition of the asymptotic lower bound.

4.2 Lower Bounds

We hereby introduce the concept of R^{RW}-consistency for the fixed-reward setting, in a manner analogous to the fixed-regret setting. Following this, we present two distinct lower bounds for the problem of regret minimization in multi-armed bandits with fixed-reward abstention in Theorem 4. The proof for Theorem 4 is postponed to Appendix E.2.

Definition 2 (R^{RW}-consistency)**.** An algorithm $\pi \in \Pi^{\mathrm{RW}}$ is said to be R^{RW}*-consistent* if for all abstention rewards $c \in \mathbb{R}$, bandit instances $\mu \in \mathcal{U}$, and $a > 0$, $R^{\mathrm{RW}}_{\mu,c}(T, \pi) = o(T^a)$.

Theorem 4. *For any abstention reward $c \in \mathbb{R}$, bandit instance $\mu \in \mathcal{U}$ and R^{RW}-consistent algorithm π, it holds that*

$$\liminf_{T\to\infty} \frac{R^{\mathrm{RW}}_{\mu,c}(T,\pi)}{\log T} \geq 2 \sum_{i>1} \frac{\mu_1 \vee c - \mu_i \vee c}{\Delta_i^2}.$$

For any abstention reward $c \in \mathbb{R}$ and time horizon $T \geq K$, there exists a universal constant $\alpha > 0$ such that

$$\inf_{\pi \in \Pi^{\mathrm{RW}}} \sup_{\mu \in \mathcal{U}} R^{\mathrm{RW}}_{\mu,c}(T,\pi) \geq \alpha\sqrt{KT}.$$

By comparing the upper bounds in Theorem 3 with the lower bounds in Theorem 4, it is firmly confirmed that Algorithm 2 is both asymptotically and minimax optimal in the fixed-reward setting.

Asymptotic optimality. For any abstention reward $c \in \mathbb{R}$ and bandit instance $\mu \in \mathcal{U}$, our algorithm ensures the following optimal asymptotic behavior for the cumulative regret:

$$\lim_{T\to\infty} \frac{R^{\mathrm{RW}}_{\mu,c}(T)}{\log T} = 2 \sum_{i>1} \frac{\mu_1 \vee c - \mu_i \vee c}{\Delta_i^2}.$$

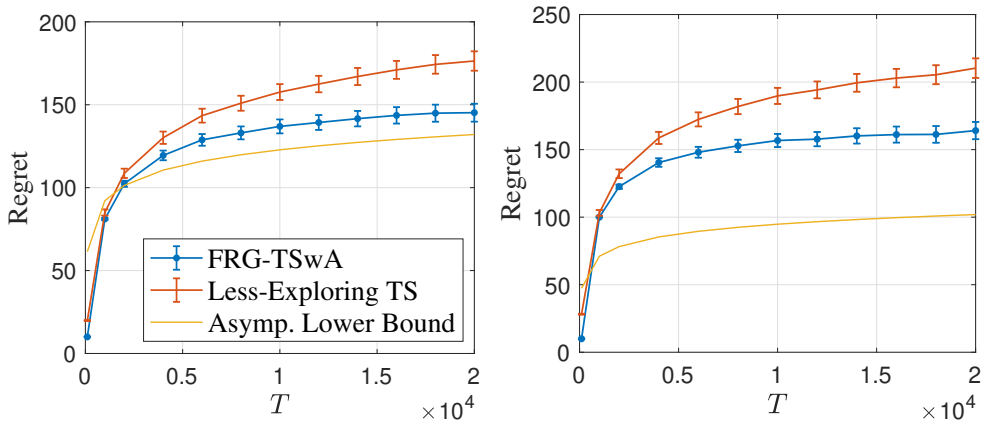
Since it holds that $\mu_1 \vee c - \mu_i \vee c \leq \Delta_i$ for all $i > 1$, our algorithm effectively reduces the cumulative regret in the asymptotic regime through the incorporation of the fixed-reward abstention option.

Minimax optimality. Concerning the worst-case performance of our algorithm, disregarding the additive term $\sum_{i\in[K]} (\mu_1 \vee c - \mu_i)$, it achieves an optimal worst-case regret of $O(\sqrt{KT})$. While this worst-case regret is the same as that for canonical multi-armed bandits, achieving it is non-trivial as we have to simultaneously achieve the minimal asymptotic instance-dependent regret.

Moreover, there is no occurrence of the phase transition phenomenon in the fixed-reward setting. This absence can be attributed to the intrinsic nature of the fixed-reward abstention option. For any abstention reward $c \in \mathbb{R}$ and online algorithm, we can always construct a challenging bandit instance that leads to a cumulative regret of $\Omega(\sqrt{KT})$, as demonstrated in the proof of the minimax lower bound in Appendix E.2.

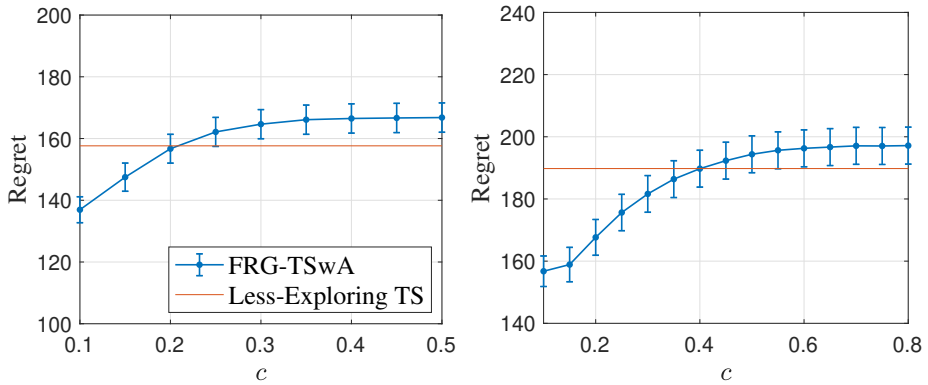
5 Numerical Experiments

In this section, we conduct numerical experiments to empirically substantiate our theoretical insights. To reduce clutter, we report our results only for the fixed-regret setting here. Results pertaining to the fixed-reward setting are deferred to Appendix F, where we consider two particular realizations of our algorithm based on Less-Exploring Thompson Sampling (Jin et al., 2023) and KL-UCB^{++} (Ménard & Garivier, 2017). In each experiment, the reported cumulative regrets are averaged over $2,000$ independent trials and the corresponding standard deviations are displayed as error bars in the figures.



(a) Instance $\mu^\dagger$ (b) Instance $\mu^\ddagger$

Figure 2: Empirical regrets with abstention regret $c = 0.1$ for different time horizons T.



(a) Instance $\mu^\dagger$ (b) Instance $\mu^\ddagger$

Figure 3: Empirical regrets with time horizon $T = 10,000$ for different abstention regrets c.

To confirm the benefits of incorporating the abstention option, we compare the performance of our proposed algorithm FRG-TSwA (Algorithm 1) with that of Less-Exploring Thompson Sampling (Jin et al., 2023), which serves as a baseline algorithm without the abstention option. We consider two synthetic bandit instances. The first instance $\mu^\dagger$ with $K = 7$ has uniform suboptimality gaps: $\mu_1^\dagger = 1$ and $\mu_i^\dagger = 0.7$ for all $i \in [K] \setminus \{1\}$. For the second instance $\mu^\ddagger$ with $K = 10$, the suboptimality gaps are more diverse: $\mu_1^\ddagger = 1$, $\mu_i^\ddagger = 0.7$ for $i \in \{2, 3, 4\}$, $\mu_i^\ddagger = 0.5$ for $i \in \{5, 6, 7\}$ and $\mu_i^\ddagger = 0.3$ for $i \in \{8, 9, 10\}$. The empirical averaged cumulative regrets of both methods with abstention regret $c = 0.1$ for different time horizons T are presented in Figure 2. To demonstrate their asymptotic behavior, we also plot the instance-dependent asymptotic lower bound on the cumulative regret (refer to Theorem 2) in each sub-figure. It can be observed that FRG-TSwA is clearly superior compared to the non-abstaining baseline, especially for large values of T. This demonstrates the advantage of the abstention mechanism. With regard to the growth trend, as the time horizon T increases, the curve corresponding to FRG-TSwA closely approximates that of the asymptotic lower bound. This suggests that the expected cumulative regret of FRG-TSwA matches the lower bound asymptotically, thereby substantiating the theoretical results presented in Section 3.

To illustrate the effect of the abstention regret c, we evaluate the performance of FRG-TSwA for varying values of c, while keeping the time horizon T fixed at $10,000$. The experimental results for both bandit instances $\mu^\dagger$ and $\mu^\ddagger$ are presented in Figure 3. Within each sub-figure, we observe that as c increases, the empirical averaged cumulative regret initially increases but eventually saturates beyond a certain threshold value of c. These empirical observations align well with our expectations. Indeed, when provided with

complete information about the bandit instance, if the abstention regret c exceeds the largest suboptimality gap, the agent gains no advantage in choosing the abstention option when selecting any arm. However, we remark that the agent lacks this oracle-like knowledge of the suboptimality gaps and must estimate them on the fly. Consequently, this results in the inevitable selection of the abstention option, even when the abstention regret c is large.

6 Conclusions and Future Work

In this paper, we consider, for the first time, a multi-armed bandit model that allows for the possibility of *abstaining* from accepting the stochastic rewards, alongside the conventional arm selection. This innovative framework is motivated by real-world scenarios where decision-makers may wish to hedge against highly uncertain or risky actions, as exemplified in clinical trials. Within this enriched paradigm, we address both the fixed-regret and fixed-reward settings, providing tight upper and lower bounds on asymptotic and minimax regrets for each scenario. For the fixed-regret setting, we thoughtfully adapt a recently developed asymptotically and minimax optimal algorithm by Jin et al. (2023) to accommodate the abstention option while preserving its attractive optimality characteristics. For the fixed-reward setting, we convert *any* asymptotically and minimax optimal algorithm for the canonical model into one that retains these optimality properties when the abstention option is present. We remark that our proposed algorithms are computationally efficient, incurring only a negligible constant-time overhead per step beyond the standard arm selection mechanisms. Finally, experiments on synthetic datasets validate our theoretical results and clearly demonstrate the advantage of incorporating the abstention option.

While our analysis focuses on Gaussian rewards, the framework naturally extends to other distributions, provided that specific boundary cases are handled with care. As highlighted in Remark 5, a fruitful avenue for future research lies in expanding the abstention model from K-armed bandits to linear bandits; specifically, it remains an open question whether the inclusion of the abstention feature can lead to enhanced asymptotic and minimax theoretical guarantees. Furthermore, in our work, the abstention option exerts no influence on the stochastic observation from the selected arm. Delving into more sophisticated and general approaches to model the effect of the abstention option promises as a captivating future direction.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. T. Jin and V. Y. F. Tan are supported by a Singapore Ministry of Education (MOE) AcRF Tier 2 grant under grant number A-8004062-00-00.

References

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory (COLT)*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012.

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for Thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6): 1785–1826, 2017.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory (COLT)*, volume 7, pp. 1–122, 2009.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pp. 322–331. IEEE, 1995.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, pp. 1516–1541, 2013.

Weiwei Cheng, Michaël Rademaker, Bernard De Baets, and Eyke Hüllermeier. Predicting partial orders: Ranking with abstention. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010*, pp. 215–230. Springer, 2010.

Chi-Keung Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, pp. 247–254, 1957.

Chi-Keung Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pp. 67–82. Springer, 2016.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online learning with abstention. In *International Conference on Machine Learning*, pp. 1059–1067. PMLR, 2018.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Leslie Meltzer Henry, Megan E Larkin, and Elizabeth R Pike. Just compensation: a no-fault proposal for research-related injuries. *Journal of Law and the Biosciences*, 2(3):645–668, 2015.

Radu Herbei and Marten H Wegkamp. Classification with reject option. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, pp. 709–721, 2006.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Conference on Learning Theory*, pp. 67–79. Citeseer, 2010.

Yu Hu, Jiwoong Shin, and Zhulei Tang. Incentive problems in performance-based online advertising pricing: Cost per click vs. cost per action. *Management Science*, 62(7):2022–2038, 2016.

Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pp. 5074–5083. PMLR, 2021.

Tianyuan Jin, Xianglin Yang, Xiaokui Xiao, and Pan Xu. Thompson sampling with less exploration is fast and optimal. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 15239–15261. PMLR, 2023.

Adam Kalai and Varun Kanade. Towards optimally abstaining from prediction with ood test examples. *Advances in Neural Information Processing Systems*, 34:12774–12785, 2021.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23*, pp. 199–213. Springer, 2012.

Nathaniel Korda, Emilie Kaufmann, and Rémi Munos. Thompson sampling for 1-dimensional exponential family bandits. *Advances in Neural Information Processing Systems*, 26, 2013.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research*, 19(1):765–796, 2018.

Tor Lattimore and Csaba Szepesvári. The end of optimism? An asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pp. 728–737. PMLR, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

Fang Liu, Swapna Buccapatnam, and Ness Shroff. Information directed sampling for stochastic bandits with graph feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Ranking with abstention. *arXiv preprint arXiv:2307.02035*, 2023.

Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pp. 223–237. PMLR, 2017.

Renuka Munshi and Urmila Thatte. Compensation for research related injury. *Perspectives in clinical research*, 4(1):61–69, 2013.

Hamid Nazerzadeh, Amin Saberi, and Rakesh Vohra. Dynamic cost-per-action mechanisms and applications to online advertising. In *Proceedings of the 17th international conference on World Wide Web*, pp. 179–188, 2008.

Gergely Neu and Nikita Zhivotovskiy. Fast rates for online prediction with abstention. In *Conference on Learning Theory*, pp. 3030–3048. PMLR, 2020.

Robert Steinbrook. Compensation for injured research subjects. *The New England journal of medicine*, 354 (18):1871–1873, 2006.

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

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

Yair Wiener and Ran El-Yaniv. Pointwise tracking the optimal regression function. *Advances in Neural Information Processing Systems*, 25, 2012.

Ahmed Zaoui, Christophe Denis, and Mohamed Hebiri. Regression with reject option and application to knn. *Advances in Neural Information Processing Systems*, 33:20073–20082, 2020.

A Asymptotic and Minimax Optimality in Canonical Multi-Armed Bandits

In the canonical multi-armed bandit model, there is no additional abstention option. Given a bandit instance $\mu \in \mathcal{U}$, at each time step $t \in \mathbb{N}$, the agent employs an *online algorithm* π to choose an arm A_t from the arm set $[K]$, and then observes a random variable X_t from the selected arm A_t, which is drawn from a Gaussian distribution $\mathcal{N}(\mu_{A_t}, 1)$ and independent of observations from previous time steps. The choice of A_t might depend on the prior decisions and observations. To describe the setup formally, A_t is $\mathcal{F}_{t-1}^{\mathrm{CA}}$-measurable, where $\mathcal{F}_t^{\mathrm{CA}} := \sigma(A_1, X_1, A_2, X_2, \ldots, A_t, X_t)$ represents the σ-field generated by the cumulative interaction history up to and including time t. Subsequently, the agent suffers an instantaneous regret of $\mu_1 - X_t$.

The agent aims at minimizing the expected cumulative regret over a time horizon T, which is defined as

$$R_\mu^{\mathrm{CA}}(T, \pi) = T\mu_1 - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right].$$

We refer to the collection of all online policies for the canonical multi-armed bandit model as Π^{CA}.

Remark 7. It is worth noting that any algorithm designed for canonical multi-armed bandit model possesses the capability to decide the arm A_t to pull at each time step t, based on the partial interaction history $(A_1, X_1, A_2, X_2, \ldots, A_{t-1}, X_{t-1})$, within the abstention model. Conversely, any algorithm tailored for the abstention model in the fixed-regret setting (or in the fixed-reward setting) can be applied to the canonical multi-armed bandit model, provided that the abstention regret (or the abstention reward) has been predetermined. Specifically, the algorithm can determine both the selected arm A_t and the binary abstention variable B_t, although B_t is purely auxiliary and exerts no influence on the cumulative regret $R_\mu^{\mathrm{CA}}(T, \pi)$.

Lower bounds. Both the asymptotic and minimax lower bounds for the canonical multi-armed bandit model have been thoroughly established (Lai & Robbins, 1985; Auer et al., 1995). For a comprehensive overview, refer to Sections 15 and 16 of Lattimore & Szepesvári (2020). Here, we summarize the results in the following:

Definition 3 (R^{CA}-consistency)**.** An algorithm $\pi \in \Pi^{\mathrm{CA}}$ is said to be R^{CA}*-consistent* if for all bandit instances $\mu \in \mathcal{U}$ and $a > 0$, $R_\mu^{\mathrm{CA}}(T, \pi) = o(T^a)$.

Theorem 5. *For any bandit instance $\mu \in \mathcal{U}$ and R^{CA}-consistent algorithm π, it holds that*

$$\liminf_{T\to\infty} \frac{R_\mu^{\mathrm{CA}}(T, \pi)}{\log T} \geq \sum_{i>1} \frac{2}{\Delta_i}.$$

For any time horizon $T \geq K$, there exists a universal constant $\alpha > 0$ such that

$$\inf_{\pi\in\Pi^{\mathrm{CA}}} \sup_{\mu\in\mathcal{U}} R_\mu^{\mathrm{CA}}(T, \pi) \geq \alpha\sqrt{KT}.$$

Asymptotic and minimax optimality. According to Theorem 5, in the canonical bandit model, an algorithm $\pi \in \Pi^{\mathrm{CA}}$ is said to be *asymptotically optimal* if for all bandit instances $\mu \in \mathcal{U}$, it ensures that

$$\lim_{T\to\infty} \frac{R_\mu^{\mathrm{CA}}(T, \pi)}{\log T} = \sum_{i>1} \frac{2}{\Delta_i}.$$

Furthermore, it is said to be *minimax optimal* if there exists a universal constant $\alpha > 0$ such that

$$R_\mu^{\mathrm{CA}}(T, \pi) \leq \alpha \left(\sqrt{KT} + \sum_{i>1} \Delta_i \right).$$

To the best of our knowledge, for canonical multi-armed bandits with Gaussian rewards, KL-UCB^{++} (Ménard & Garivier, 2017), ADA-UCB (Lattimore, 2018), MOTS-$\mathcal{J}$ (Jin et al., 2021) and Less-Exploring Thompson Sampling (Jin et al., 2023) exhibit simultaneous asymptotic and minimax optimality. As their names suggest, the former two algorithms follow the UCB-style, while the latter two are rooted in Thompson Sampling.

Remark 8. One valuable byproduct derived from the proof of the asymptotic lower bound in Theorem 5 is that, for any R^{CA}-consistent algorithm π, bandit instance $\mu \in \mathcal{U}$ and suboptimal arm $i > 1$, we have

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_i(T)]}{\log T} \geq \frac{2}{\Delta_i^2}.$$

Therefore, any algorithm that is asymptotically optimal ensures that for all suboptimal arms $i > 1$,

$$\lim_{T\to\infty} \frac{\mathbb{E}[N_i(T)]}{\log T} = \frac{2}{\Delta_i^2}.$$

B Comparison with the $(K+1)$-Armed Bandit Reduction

A natural question arises regarding the formulation of our abstention model: can the problem be simply reduced to a standard multi-armed bandit setting with $K+1$ arms? In such a reduction, the additional $(K+1)$-st arm would represent the abstention option, yielding a deterministic reward c (in the fixed-reward setting) or a fixed regret (in the fixed-regret setting).

While such a reduction might appear intuitive and indeed may suffice for achieving coarse worst-case (minimax) guarantees of $O(\sqrt{KT})$, it fundamentally fails to capture the strategic advantages of our framework. Specifically, the reduction cannot achieve the same instance-dependent asymptotic optimality. In the following, we discuss the essential differences between our model and the $(K+1)$-armed reduction.

Structural Difference: Decoupling Learning from Reward. The most critical distinction lies in the mechanism of information acquisition. In a standard $(K+1)$-armed bandit model, selecting the deterministic $(K+1)$-th arm (the "safe" option) fixes the instantaneous outcome but reveals *no information* regarding the distributions of the original K stochastic arms. To learn about a specific arm $i \in [K]$, the agent is forced to pull it and fully accept its stochastic outcome.

In contrast, our abstention framework allows the agent to *decouple* the observation of an outcome from the acceptance of its reward. When the agent exercises the abstention option, they secure a guaranteed result (equivalent to the deterministic arm) but, crucially, *simultaneously* observe the stochastic sample X_t from the selected arm A_t. This structure enables what can be termed "safe exploration": the agent can continue to gather valuable statistics about potentially risky or unknown arms to refine their estimates, all while shielding themselves from the full cost of stochastic underperformance.

Performance Gap: Asymptotic Optimality. To illustrate how this structural difference translates into superior asymptotic performance, we show below that in both the fixed-reward and fixed-regret objectives, the $(K+1)$-armed reduction and our abstention model admit different instance-dependent asymptotic behavior.

Fixed Reward: In a standard $(K+1)$-armed bandit reduction, to learn about a suboptimal arm i, the agent must select it. Consequently, the agent incurs a regret proportional to the full suboptimality gap $\Delta_i = \mu_1 - \mu_i$ for every exploratory sample. The asymptotic regret contribution from arm i is thus driven by the factor Δ_i. However, in our fixed-reward abstention model, the agent can leverage the abstention option to mitigate this cost. Specifically, if the abstention reward c satisfies $\mu_i < c < \mu_1$, the agent can pull arm i while simultaneously choosing to abstain. In this scenario, the agent observes the outcome of arm i (gaining information) but receives the fixed reward c. The resulting instantaneous regret is $\mu_1 - c$, which is strictly less than the full gap $\mu_1 - \mu_i$.

This mechanism facilitates a more cost-effective form of exploration. Consequently, our derived asymptotic regret bound scales with the term $\mu_1 \vee c - \mu_i \vee c$, rather than the larger gap Δ_i that arises from the (K+1)-armed reduction. This confirms that a simple reduction to a $(K+1)$-armed bandit cannot replicate the superior asymptotic optimality achieved by our proposed model.

Fixed Regret: A simple $(K+1)$-armed reduction also fails in the fixed-regret objective. In that reduction, the $(K+1)$-st "abstain" arm incurs a deterministic regret c, but it reveals *no information* about the original K stochastic arms. Hence, to statistically verify that a suboptimal arm i is worse than the best arm, the learner must still pull arm i and *accept* its stochastic outcome, paying regret about Δ_i per sample. By the standard asymptotic lower bound in Lai & Robbins (1985), the regret caused by pulling arm i is on the order of $\frac{2\log T}{\Delta_i}$. Therefore, the leading asymptotic term under the $(K+1)$-arm reduction is $\sum_{i>1} \frac{2\log T}{\Delta_i}$, which is *independent of* c. In contrast, in our abstention model the learner can pull arm i while abstaining, still observing X_t but capping the instantaneous regret by c. Thus, exploration of arm i costs $\Delta_i \wedge c$ per sample, yielding the strictly smaller scaling $2\sum_{i>1} \frac{\Delta_i \wedge c}{\Delta_i^2} \log T$, and in particular improves $\frac{2\log T}{\Delta_i}$ to $\frac{2c\log T}{\Delta_i^2}$ whenever $\Delta_i > c$. This again shows the $(K+1)$-armed reduction cannot replicate the fixed-regret asymptotic achieved by our model.

C Auxiliary Lemmas

Lemma 1 (Bretagnolle–Huber inequality (Tsybakov, 2009))**.** *Let $\mathbb{P}$ and $\mathbb{P}'$ be two probability distributions on the same measurable space $(\Omega, \mathcal{F})$. For any event $A \in \mathcal{F}$ and its complement $A^c = \Omega \setminus A$, the following inequality holds:*

$$\mathbb{P}(A) + \mathbb{P}'(A^c) \geq \frac{1}{2}\exp(-\mathrm{KL}(\mathbb{P}, \mathbb{P}')),$$

where $\mathrm{KL}(\mathbb{P}, \mathbb{P}')$ denotes the Kullback–Leibler (KL) divergence between $\mathbb{P}$ and $\mathbb{P}'$.

Lemma 2 (Divergence decomposition lemma)**.** *Consider both the fixed-regret setting and the fixed-reward setting. Fix an arbitrary policy π. Let $\nu = (\mathbb{P}_1, \ldots, \mathbb{P}_K)$ represent the reward distributions associated with one bandit instance, and let $\nu' = (\mathbb{P}'_1, \ldots, \mathbb{P}'_K)$ represent the reward distributions associated with another bandit instance. Define $\mathbb{P}_{\nu,c}$ as the probability distribution of the sequence $(A_1, B_1, X_1, \ldots, A_T, B_T, X_T)$ induced by the algorithm π under the abstention regret c in the fixed-regret setting (or the abstention reward c in the fixed-reward setting) for the bandit instance ν. Similarly, let $\mathbb{P}_{\nu',c}$ denote the same for the bandit instance ν'. Then the KL divergence between $\mathbb{P}_{\nu,c}$ and $\mathbb{P}_{\nu',c}$ can be decomposed as:*

$$\mathrm{KL}\left(\mathbb{P}_{\nu,c}, \mathbb{P}_{\nu',c}\right) = \sum_{i\in[K]} \mathbb{E}_{\nu,c}[N_i(T)]\mathrm{KL}\left(\mathbb{P}_i, \mathbb{P}'_i\right).$$

The proof of Lemma 2 is similar to the well-known proof of divergence decomposition in the canonical multi-armed bandit model (excluding abstention), and is therefore omitted. This proof can be located, for instance, in Garivier et al. (2019, Section 2.1) and Lattimore & Szepesvári (2020, Lemma 15.1).

Lemma 3 (Hoeffding's inequality for sub-Gaussian random variables)**.** *Let $X_1, \ldots, X_n$ be independent σ-sub-Gaussian random variables with mean μ. Then for any $\varepsilon \geq 0$,*

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) \quad \textit{and} \quad \mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$$

where $\hat{\mu} := \frac{1}{n}\sum_{i=1}^n X_i$.

Lemma 4. *Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of independent σ-sub-Gaussian random variables with mean μ. Then for any $\varepsilon > 0$ and $N \in \mathbb{N}$,*

$$\sum_{n=1}^N \mathbb{P}(\hat{\mu}_n \geq \mu + \varepsilon) \leq \frac{2\sigma^2}{\varepsilon^2} \quad \textit{and} \quad \sum_{n=1}^N \mathbb{P}(\hat{\mu}_n \leq \mu - \varepsilon) \leq \frac{2\sigma^2}{\varepsilon^2}$$

where $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n X_i$.

Proof. By symmetry, it suffices to prove the first part. According to Lemma 3, we have

$$\begin{aligned}
\sum_{n=1}^N \mathbb{P}(\hat{\mu}_n \geq \mu + \varepsilon) &\leq \sum_{n=1}^N \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) \\
&\leq \frac{\exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)}{1 - \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)} \\
&= \frac{1}{\exp\left(\frac{\varepsilon^2}{2\sigma^2}\right) - 1} \\
&\leq \frac{2\sigma^2}{\varepsilon^2}
\end{aligned}$$

where the last inequality follows from the fact that $e^x - 1 \geq x$ for any $x \geq 0$. □

D Analysis of the Fixed-Regret Setting

D.1 Upper Bounds

Proof of Theorem 1. Due to the law of total expectation, we can decompose the regret $R^{\mathrm{RG}}_{\mu,c}(T,\pi)$ as

$$\begin{aligned} R^{\mathrm{RG}}_{\mu,c}(T,\pi) &= \mathbb{E}\left[\sum_{t=1}^{T}\left((\mu_1 - X_t)\cdot \mathbf{1}\{B_t=0\} + c\cdot \mathbf{1}\{B_t=1\}\right)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\left((\mu_1 - \mu_{A_t})\cdot \mathbf{1}\{B_t=0\} + c\cdot \mathbf{1}\{B_t=1\}\right)\right] \\ &= c\cdot \mathbb{E}[N_1^{(1)}(T)] + \sum_{i>1}\left(\Delta_i\cdot \mathbb{E}[N_i^{(0)}(T)] + c\cdot \mathbb{E}[N_i^{(1)}(T)]\right). \end{aligned} \tag{4}$$

For any arm i with $\Delta_i < c$ (including the best arm), it holds that

$$\begin{aligned} &\mathbb{E}[N_i^{(1)}(T)] \\ =& \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{A_t = i \text{ and } B_t = 1\}\right] \\ \le& \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{A_t = i \text{ and } \sqrt{\frac{K}{t}} \ge c\right\}\right] \\ &+ \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{I}\left\{A_t = i \text{ and } \max_{j\in[K]\setminus\{i\}}\left(\hat{\mu}_j(t-1) - \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_j(t-1)}}\right) - \hat{\mu}_i(t-1) \ge c\right\}\right] \\ \le& \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{A_t = i \text{ and } \sqrt{\frac{K}{t}} \ge c\right\}\right] \\ &+ \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{I}\left\{A_t = i \text{ and } \max_{j\in[K]}\left(\hat{\mu}_j(t-1) - \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_j(t-1)}}\right) \ge \mu_1\right\}\right] \\ &+ \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{I}\{A_t = i \text{ and } \mu_1 - \hat{\mu}_i(t-1) \ge c\}\right] \end{aligned} \tag{5}$$

where the last inequality arises from the observation that when $A_t = i$,

$$\begin{aligned} &\left\{\max_{j\in[K]\setminus\{i\}}\left(\hat{\mu}_j(t-1) - \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_j(t-1)}}\right) - \hat{\mu}_i(t-1) \ge c\right\} \\ \subseteq& \left\{\max_{j\in[K]\setminus\{i\}}\left(\hat{\mu}_j(t-1) - \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_j(t-1)}}\right) \ge \mu_1\right\} \cup \{\mu_1 - \hat{\mu}_i(t-1) \ge c\} \\ \subseteq& \left\{\max_{j\in[K]}\left(\hat{\mu}_j(t-1) - \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_j(t-1)}}\right) \ge \mu_1\right\} \cup \{\mu_1 - \hat{\mu}_i(t-1) \ge c\}. \end{aligned}$$

For convenience, for any $i \in [K]$ such that $\Delta_i < c$, we introduce three shorthand notations to represent the terms in Equation (5). These terms correspond to distinct conditions under which the algorithm chooses to abstain:

$$\begin{cases} (\clubsuit)_i := \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{A_t = i \text{ and } \sqrt{\frac{K}{t}} \ge c\right\}\right] \\ (\spadesuit)_i := \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{I}\left\{A_t = i \text{ and } \max_{j\in[K]}\left(\hat{\mu}_j(t-1) - \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_j(t-1)}}\right) \ge \mu_1\right\}\right] \\ (\blacksquare)_i := \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{I}\{A_t = i \text{ and } \mu_1 - \hat{\mu}_i(t-1) \ge c\}\right]. \end{cases}$$

To clarify the intuition behind these terms:

- $(\clubsuit)_i$ represents the gap-independent abstention. This captures the case where the algorithm abstains purely because the time step t is small, independent of the empirical estimates.
- $(\spadesuit)_i$ represents abstentions caused by the overestimation of LCBs. This captures the low-probability event where the lower confidence bound of any arm j exceeds the true mean of the optimal arm μ_1.
- $(\blacksquare)_i$ represents abstentions caused by the underestimation of the selected arm. This captures the event where the empirical mean of the chosen arm i deviates significantly below its true mean μ_i (specifically such that $\mu_1 - \hat{\mu}_i(t-1) \geq c$), triggering the gap-dependent abstention criterion.

We will analyze $(\clubsuit)_i$ and $(\spadesuit)_i$ later to establish the two forms of upper bounds.

On the other hand, for the term $(\blacksquare)_i$, which relates strictly to the concentration of arm i, we have

$$\begin{aligned}
(\blacksquare)_i &\leq \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{s=1}^{T-1}\mathbb{I}\left\{A_t = i \text{ and } \hat{\mu}_{is} \leq \mu_i - (c-\Delta_i) \text{ and } N_i(t-1) = s\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{s=1}^{T-1}\mathbb{I}\left\{\hat{\mu}_{is} \leq \mu_i - (c-\Delta_i)\right\}\right] \qquad (6) \\
&\leq \frac{2}{(c-\Delta_i)^2}. \qquad (7)
\end{aligned}$$

Line (6) follows from the fact that for all $s \in [T-1]$,

$$\sum_{t=K+1}^{T}\mathbb{I}\left\{A_t = i \text{ and } N_i(t-1) = s\right\} \leq 1.$$

Line (7) is due to Lemma 4.

For any arm i with $\Delta_i > c$, since arm $1 \in [K] \setminus \{i\}$, we have

$$\begin{aligned}
&\mathbb{E}[N_i^{(0)}(T)] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{A_t = i \text{ and } B_t = 0\right\}\right] \\
&\leq 1 + \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{I}\left\{A_t = i \text{ and } \right.\right. \\
&\qquad\qquad \left.\left. \max_{j\in[K]\setminus\{i\}}\left(\hat{\mu}_j(t-1) - \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_j(t-1)}}\right) - \hat{\mu}_i(t-1) < c \text{ and } \sqrt{\frac{K}{t}} < c\right\}\right] \\
&\leq 1 + \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{I}\left\{A_t = i \text{ and } \left(\hat{\mu}_1(t-1) - \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_1(t-1)}}\right) - \hat{\mu}_i(t-1) < c\right\}\right]. \qquad (8)
\end{aligned}$$

Minimax upper bound. If $c \leq \sqrt{\frac{K}{T}}$, then the abstention option is always invoked because $\sqrt{\frac{K}{t}} \geq \sqrt{\frac{K}{T}} \geq c$ for all $t \in [T]$. Consequently, it is straightforward to deduce that

$$R^{\mathrm{RG}}_{\mu,c}(T) \leq cT.$$

Next, consider the case that $c > \sqrt{\frac{K}{T}}$. Compared with the canonical multi-armed bandit model, at a single time step, the agent in our abstention model incurs a greater (expected) regret only if an arm i with $\Delta_i < c$ is pulled and the abstention option is selected. Thus, we have

$$R^{\mathrm{RG}}_{\mu,c}(T) \leq R^{\mathrm{CA}}_{\mu}(T) + \sum_{i:\Delta_i<c} (c - \Delta_i) \cdot \mathbb{E}[N_i^{(1)}(T)]. \tag{9}$$

Due to the minimax optimality of Less-Exploring Thompson Sampling (Jin et al., 2023), there exists a universal constant $\alpha_1 > 0$ such that

$$R^{\mathrm{CA}}_{\mu}(T) \leq \alpha_1 \left(\sqrt{KT} + \sum_{i>1} \Delta_i \right). \tag{10}$$

Recall the upper bound of $\mathbb{E}[N_i^{(1)}(T)]$ for arm i with $\Delta_i < c$, as given in (5). Subsequently, we will establish bounds for the following terms:

$$\sum_{i:\Delta_i<c} (c - \Delta_i) \cdot (\clubsuit)_i \ , \quad \sum_{i:\Delta_i<c} (c - \Delta_i) \cdot (\spadesuit)_i \quad \text{and} \quad \sum_{i:\Delta_i<c} (c - \Delta_i) \cdot ((\blacksquare)_i.$$

For the first term, we have

$$\begin{aligned}
\sum_{i:\Delta_i<c} (c - \Delta_i) \cdot (\clubsuit)_i &\leq \sum_{i:\Delta_i<c} c \cdot \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left\{ A_t = i \text{ and } \sqrt{\frac{K}{t}} \geq c \right\} \right] \\
&\leq c \cdot \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left\{ \sqrt{\frac{K}{t}} \geq c \right\} \right] \\
&\leq \frac{K}{c} \\
&\leq \sqrt{KT}.
\end{aligned} \tag{11}$$

For the second term, we can obtain

$$\begin{aligned}
\sum_{i:\Delta_i<c} (c - \Delta_i) \cdot (\spadesuit)_i &\leq \sum_{i:\Delta_i<c} c \cdot (\spadesuit)_i \\
&\leq c \cdot \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{ \max_{j\in[K]} \left(\hat{\mu}_j(t-1) - \sqrt{\frac{6\log t + 2\log(c \vee 1)}{N_j(t-1)}} \right) \geq \mu_1 \right\} \right] \\
&\leq c \cdot \sum_{j\in[K]} \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{ \hat{\mu}_j(t-1) - \sqrt{\frac{6\log t + 2\log(c \vee 1)}{N_j(t-1)}} \geq \mu_1 \right\} \right].
\end{aligned}$$

For all $j \in [K]$, by a union bound over all possible values of $N_j(t-1)$ and Lemma 3, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{\hat{\mu}_j(t-1) - \sqrt{\frac{6\log t + 2\log(c \vee 1)}{N_j(t-1)}} \geq \mu_1\right\}\right] \\
&\leq \sum_{t=K+1}^{T}\sum_{s=1}^{t-1} \mathbb{P}\left(\hat{\mu}_{js} - \sqrt{\frac{6\log t + 2\log(c \vee 1)}{s}} \geq \mu_1\right) \\
&\leq \sum_{t=K+1}^{T}\sum_{s=1}^{t-1} \mathbb{P}\left(\hat{\mu}_{js} - \sqrt{\frac{6\log t + 2\log(c \vee 1)}{s}} \geq \mu_j\right) \\
&\leq \sum_{t=K+1}^{T}\sum_{s=1}^{t-1} \frac{1}{t^3(c \vee 1)} \\
&\leq \sum_{t=K+1}^{T} \frac{1}{t^2 c} \\
&\leq \frac{1}{Kc}
\end{aligned} \tag{12}$$

where the last inequality follows from the numerical fact that

$$\sum_{t=K+1}^{T} \frac{1}{t^2} \leq \int_{x=K}^{\infty} \frac{1}{x^2}\,\mathrm{d}x = \frac{1}{K}.$$

Thus, we can bound the second term as

$$\sum_{i:\Delta_i<c} (c-\Delta_i)\cdot(\spadesuit)_i \leq c \cdot \sum_{j\in[K]} \frac{1}{Kc} = 1. \tag{13}$$

For the third term, in addition to the upper bound of $(\blacksquare)_i$ in (7), we can identify another straightforward upper bound as follows:

$$(\blacksquare)_i \leq \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\{A_t = i\}\right] \leq \mathbb{E}[N_i(T)].$$

By applying these two bounds separately for distinct scenarios,, we have

$$\begin{aligned}
\sum_{i:\Delta_i<c} (c-\Delta_i)\cdot(\blacksquare)_i &\leq \sum_{i:0<c-\Delta_i<\sqrt{\frac{K}{T}}} (c-\Delta_i)\cdot(\blacksquare)_i + \sum_{i:c-\Delta_i\geq\sqrt{\frac{K}{T}}} (c-\Delta_i)\cdot(\blacksquare)_i \\
&\leq \sqrt{\frac{K}{T}} \sum_{i:0<c-\Delta_i<\sqrt{\frac{K}{T}}} \mathbb{E}[N_i(T)] + \sum_{i:c-\Delta_i\geq\sqrt{\frac{K}{T}}} \frac{2}{c-\Delta_i} \\
&\leq \sqrt{\frac{K}{T}}\cdot T + \sum_{i:c-\Delta_i\geq\sqrt{\frac{K}{T}}} 2\sqrt{\frac{T}{K}} \\
&\leq \sqrt{KT} + K\cdot 2\sqrt{\frac{T}{K}} \\
&= 3\sqrt{KT}.
\end{aligned} \tag{14}$$

By plugging Inequalities (11), (13) and (14) into (5), we have

$$\sum_{i:\Delta_i<c} (c-\Delta_i)\cdot\mathbb{E}[N_i^{(1)}(T)] \leq 1 + 4\sqrt{KT}.$$

Together with (9) and (10), we can conclude that

$$R_{\mu,c}^{\mathrm{RG}}(T) \leq (\alpha_1 + 4)\sqrt{KT} + \alpha_1 \sum_{i>1} \Delta_i + 1.$$

Therefore, there must exist a universal constant $\alpha > 0$ such that

$$R_{\mu,c}^{\mathrm{RG}}(T) \leq \alpha \left(\sqrt{KT} + \sum_{i>1} \Delta_i \right).$$

This completes the proof of the minimax upper bound.

Asymptotic upper bound. Consider any arm i with $\Delta_i < c$ (including the best arm). In the following, we will further elucidate the upper bound of $\mathbb{E}[N_i^{(1)}(T)]$ as given in (5) within the asymptotic domain.

For the first term $(\clubsuit)_i$ in (5), we have

$$(\clubsuit)_i \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left\{ \sqrt{\frac{K}{t}} \geq c \right\} \right] \leq \frac{K}{c^2}.$$

For the second term $(\spadesuit)_i$, using Inequality (12), we can get

$$\begin{aligned}
(\spadesuit)_i &\leq \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{ \max_{j\in[K]} \left(\hat{\mu}_j(t-1) - \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_j(t-1)}} \right) \geq \mu_1 \right\} \right] \\
&\leq \sum_{j\in[K]} \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{ \hat{\mu}_j(t-1) - \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_j(t-1)}} \geq \mu_1 \right\} \right] \\
&\leq \frac{1}{c}.
\end{aligned}$$

Incorporating (7), we obtain

$$\mathbb{E}[N_i^{(1)}(T)] \leq \frac{K}{c^2} + \frac{1}{c} + \frac{2}{(c-\Delta_i)^2} = o(\log T).$$

Consider any arm i with $\Delta_i > c$. We will further explore the upper bound of $\mathbb{E}[N_i^{(0)}(T)]$ in (8).

According to the fact that

$$\begin{aligned}
&\left\{ A_t = i \text{ and } \left(\hat{\mu}_1(t-1) - \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_1(t-1)}} \right) - \hat{\mu}_i(t-1) < c \right\} \\
\subseteq &\left\{ \hat{\mu}_1(t-1) + \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_1(t-1)}} \leq \mu_1 \right\} \\
&\cup \left\{ A_t = i \text{ and } \left(\mu_1 - 2\sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_1(t-1)}} \right) - \hat{\mu}_i(t-1) \leq c \right\},
\end{aligned}$$

we have

$$\begin{aligned}
&\mathbb{E}[N_i^{(0)}(T)] \\
\leq &1 + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{ A_t = i \text{ and } \left(\hat{\mu}_1(t-1) - \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_1(t-1)}} \right) - \hat{\mu}_i(t-1) < c \right\} \right]
\end{aligned}$$

$$
\begin{aligned}
&\leq 1 + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{\hat{\mu}_1(t-1) + \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_1(t-1)}} \leq \mu_1\right\}\right] \\
&\quad + \underbrace{\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{A_t = i \text{ and } \left(\mu_1 - 2\sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_1(t-1)}}\right) - \hat{\mu}_i(t-1) < c\right\}\right]}_{(\bigstar)_i}.
\end{aligned} \tag{15}
$$

Following a similar argument as in (12), we can derive

$$
\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{\hat{\mu}_1(t-1) + \sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_1(t-1)}} \leq \mu_1\right\}\right] \leq \frac{1}{Kc}. \tag{16}
$$

Now we focus on the last term in (15), which is denoted by $(\bigstar)_i$.

For any fixed $b \in (0,1)$, there must exist a constant $t_1(b,\mu,c) \geq K+1$ such that for all $t \geq t_1$,

$$
2\sqrt{\frac{6\log t + 2\log(c\vee 1)}{(t-1)^b}} \leq \frac{\Delta_i - c}{2}.
$$

Notice that for all $t \geq t_1$,

$$
\begin{aligned}
&\left\{A_t = i \text{ and } \left(\mu_1 - 2\sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_1(t-1)}}\right) - \hat{\mu}_i(t-1) < c\right\} \\
&\subseteq \left\{N_1(t-1) \leq (t-1)^b\right\} \\
&\quad \cup \left\{A_t = i \text{ and } \left(\mu_1 - 2\sqrt{\frac{6\log t + 2\log(c\vee 1)}{N_1(t-1)}}\right) - \hat{\mu}_i(t-1) < c \text{ and } N_1(t-1) > (t-1)^b\right\} \\
&\subseteq \left\{N_1(t-1) \leq (t-1)^b\right\} \cup \left\{A_t = i \text{ and } \hat{\mu}_i(t-1) \geq \mu_1 + \frac{\Delta_i - c}{2}\right\}.
\end{aligned}
$$

From the above, we deduce that

$$
(\bigstar)_i \leq t_1 + \sum_{t=t_1}^{T} \mathbb{P}\left(N_1(t-1) \leq (t-1)^b\right) + \mathbb{E}\left[\sum_{t=t_1}^{T} \mathbb{I}\left\{A_t = i \text{ and } \hat{\mu}_i(t-1) \geq \mu_1 + \frac{\Delta_i - c}{2}\right\}\right].
$$

Using the approach similar to the one used to bound $(\blacksquare)_i$ in (7), we have

$$
\mathbb{E}\left[\sum_{t=t_1}^{T} \mathbb{I}\left\{A_t = i \text{ and } \hat{\mu}_i(t-1) \geq \mu_1 + \frac{\Delta_i - c}{2}\right\}\right] \leq \frac{8}{(\Delta_i - c)^2}.
$$

By applying Lemma 5, we can get

$$
(\bigstar)_i \leq t_1(b,\mu,c) + \beta(b,\mu,K) + \frac{8}{(\Delta_i - c)^2}
$$

where the term $\beta(b,\mu,K)$ is subsequently defined in Lemma 5.

Substituting the above inequality and (16) into (15), we arrive at

$$
\mathbb{E}[N_i^{(0)}(T)] \leq 1 + \frac{1}{Kc} + t_1(b,\mu,c) + \beta(b,\mu,K) + \frac{8}{(\Delta_i - c)^2}
$$

$$= o(\log T).$$

Due to the asymptotic optimality of Less-Exploring Thompson Sampling (Jin et al., 2023), for any suboptimal arm i, we have

$$\mathbb{E}[N_i(T)] \leq \frac{2\log T}{\Delta_i^2} + o(\log T).$$

Finally, based on the regret decomposition in (4), we can conclude

$$\begin{aligned}
R^{\mathrm{RG}}_{\mu,c}(T) &= c \cdot \mathbb{E}[N_1^{(1)}(T)] + \sum_{i>1} \Big(\Delta_i \cdot \mathbb{E}[N_i^{(0)}(T)] + c \cdot \mathbb{E}[N_i^{(1)}(T)] \Big) \\
&= c \cdot \mathbb{E}[N_1^{(1)}(T)] + \sum_{i>1} (\Delta_i \wedge c) \cdot \mathbb{E}[N_i(T)] \\
&\quad + \sum_{i>1} \Big((\Delta_i - \Delta_i \wedge c) \cdot \mathbb{E}[N_i^{(0)}(T)] + (c - \Delta_i \wedge c) \cdot \mathbb{E}[N_i^{(1)}(T)] \Big) \\
&= \sum_{i>1} (\Delta_i \wedge c) \cdot \mathbb{E}[N_i(T)] + \sum_{i:\Delta_i<c} (c - \Delta_i) \cdot \mathbb{E}[N_i^{(1)}(T)] + \sum_{i:\Delta_i>c} (\Delta_i - c) \cdot \mathbb{E}[N_i^{(0)}(T)] \\
&\leq (2\log T) \sum_{i>1} \frac{\Delta_i \wedge c}{\Delta_i^2} + o(\log T)
\end{aligned}$$

where the second equality is due to the fact that $\mathbb{E}[N_i^{(0)}(T)] + \mathbb{E}[N_i^{(1)}(T)] = \mathbb{E}[N_i(T)]$ for all arms $i \in [K]$.

Therefore, it holds that

$$\limsup_{T\to\infty} \frac{R^{\mathrm{RG}}_{\mu,c}(T)}{\log T} \leq 2 \sum_{i>1} \frac{\Delta_i \wedge c}{\Delta_i^2}$$

as desired. □

Lemma 5. *Consider Algorithm 1. For any $b \in (0,1)$, there exists a constant $\beta(b,\mu,K)$ such that*

$$\sum_{t=1}^{\infty} \mathbb{P}\left(N_1(t) \leq t^b\right) \leq \beta(b,\mu,K).$$

Proof. The proof of Lemma 5 closely follows that of Proposition 5 in Korda et al. (2013), which was used to analyze the classical Thompson Sampling algorithm. In fact, the only difference between our arm sampling rule, which is built upon Less-Exploring Thompson Sampling (Jin et al., 2023), and the classical Thompson Sampling is how the estimated reward $a_i(t)$ is constructed for each arm $i \in [K]$. Specifically, in our arm sampling rule, $a_i(t)$ is either drawn from the posterior distribution $\mathcal{N}(\hat{\mu}_i(t-1), 1/N_i(t-1))$ with probability $1/K$ or set to be the empirical mean $\hat{\mu}_i(t-1)$ otherwise. In classical Thompson Sampling, $a_i(t)$ is always drawn from the posterior distribution. Therefore, it suffices to verify the parts concerning the probability distributions of $a_i(t)$; these correspond to Lemmas 9 and 10 in the proof of Proposition 5 in Korda et al. (2013).

It is straightforward to see that Lemma 9 in Korda et al. (2013) is applicable to our algorithm. For Lemma 10 therein, its counterpart is demonstrated in Lemma 6 below.

After establishing the counterparts of Lemmas 9 and 10 in the proof of Proposition 5 in Korda et al. (2013), we can extend the same analysis to our specific case. For the sake of completeness, we provide a proof sketch in the following.

Let τ_j denote the time of the j-th pull of the optimal arm (i.e., arm 1), with $\tau_0 := 0$. Define $\xi_j := (\tau_{j+1} - 1) - \tau_j$ as the random variable measuring the number of time steps between the j-th and $(j+1)$-th pull of the optimal arm. With this setup, we can derive an upper bound for $\mathbb{P}\left(N_1(t) \leq t^b\right)$ as:

$$\mathbb{P}\left(N_1(t) \leq t^b\right) \leq \mathbb{P}\left(\exists j \in \left\{0,..,\lfloor t^b \rfloor\right\} : \xi_j \geq t^{1-b} - 1\right) \leq \sum_{j=0}^{\lfloor t^b \rfloor} \mathbb{P}(\xi_j \geq t^{1-b} - 1).$$

Consider the interval $\mathcal{I}_j := \left\{\tau_j, \ldots, \tau_j + \left\lceil t^{1-b} - 1 \right\rceil\right\}$. If $\xi_j \geq t^{1-b} - 1$, then no pull of the optimal arm occurs on $\mathcal{I}_j$.

The subsequent analysis aims to bound the probability that no pull of the optimal arm occurs within the interval $\mathcal{I}_j$. It relies on two key principles:

- First, for a suboptimal arm, if it has been pulled a sufficient number of times, then, with high probability, its estimated reward (sample) cannot deviate significantly from its true mean. This observation is quantitatively characterized in Lemma 10 of Korda et al. (2013), corresponding to Lemma 6 in our paper.
- Second, for the optimal arm, the probability that its estimated reward (sample) deviates significantly below its true mean during a long subinterval of $\mathcal{I}_j$ is low. This observation is quantitatively characterized in Lemma 9 of Korda et al. (2013), which directly applies to our case.

□

Lemma 6. *Consider Algorithm 1. For all $t \in \mathbb{N}$, it holds that*

$$\mathbb{P}\left(\exists\, s \leq t, \exists\, i > 1 : a_i(s) > \mu_i + \frac{\Delta_i}{2}, N_i(s-1) > \frac{128 \log t}{\Delta_i^2}\right) \leq \frac{K}{t^2}.$$

Proof. For any fixed $s \leq t$ and $i > 1$, we have

$$\begin{aligned}
&\mathbb{P}\left(a_i(s) > \mu_i + \frac{\Delta_i}{2}, N_i(s-1) > \frac{128 \log t}{\Delta_i^2}\right) \\
\leq\ &\mathbb{P}\left(\hat{\mu}_i(s-1) > \mu_i + \frac{\Delta_i}{4}, N_i(s-1) > \frac{128 \log t}{\Delta_i^2}\right) \\
&+ \mathbb{P}\left(a_i(s) > \mu_i + \frac{\Delta_i}{2}, N_i(s-1) > \frac{128 \log t}{\Delta_i^2}, \hat{\mu}_i(s-1) \leq \mu_i + \frac{\Delta_i}{4}\right).
\end{aligned} \tag{17}$$

For the first term in (17), by Lemma 3, we can bound it as

$$\begin{aligned}
&\mathbb{P}\left(\hat{\mu}_i(s-1) > \mu_i + \frac{\Delta_i}{4}, N_i(s-1) > \frac{128 \log t}{\Delta_i^2}\right) \\
\leq\ &\sum_{x=\left\lceil \frac{128 \log t}{\Delta_i^2} \right\rceil}^{t} \mathbb{P}\left(\hat{\mu}_{ix} > \mu_i + \frac{\Delta_i}{4}, N_i(s-1) = x\right) \\
\leq\ &\sum_{x=\left\lceil \frac{128 \log t}{\Delta_i^2} \right\rceil}^{t} \mathbb{P}\left(\hat{\mu}_{ix} > \mu_i + \frac{\Delta_i}{4}\right) \\
\leq\ &\sum_{x=\left\lceil \frac{128 \log t}{\Delta_i^2} \right\rceil}^{t} \frac{1}{t^4} \\
\leq\ &\frac{1}{t^3}.
\end{aligned}$$

For the second term in (17), according to the construction of $a_i(s)$ in Algorithm 1 and Lemma 3, we can get

$$\mathbb{P}\left(a_i(s) > \mu_i + \frac{\Delta_i}{2}, N_i(s-1) > \frac{128 \log t}{\Delta_i^2}, \hat{\mu}_i(s-1) \leq \mu_i + \frac{\Delta_i}{4}\right)$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(a_i(s) > \hat{\mu}_i(s-1) + \frac{\Delta_i}{4}, N_i(s-1) > \frac{128\log t}{\Delta_i^2}\right) \\
&\leq \sum_{x=\left\lceil \frac{128\log t}{\Delta_i^2}\right\rceil}^{t} \mathbb{P}\left(a_i(s) > \hat{\mu}_i(s-1) + \frac{\Delta_i}{4}, N_i(s-1) = x\right) \\
&\leq \sum_{x=\left\lceil \frac{128\log t}{\Delta_i^2}\right\rceil}^{t} \frac{1}{K}\cdot\frac{1}{t^4} \\
&\leq \frac{1}{Kt^3}.
\end{aligned}
$$

Thus, for any $s \leq t$ and $i > 1$, it holds that

$$
\mathbb{P}\left(a_i(s) > \mu_i + \frac{\Delta_i}{2}, N_i(s-1) > \frac{128\log t}{\Delta_i^2}\right) \leq \frac{1}{t^3} + \frac{1}{Kt^3} = \frac{K+1}{Kt^3}.
$$

Finally, by a union bound, we can conclude

$$
\mathbb{P}\left(\exists\, s \leq t, \exists\, i > 1 : a_i(s) > \mu_i + \frac{\Delta_i}{2}, N_i(s-1) > \frac{128\log t}{\Delta_i^2}\right) \leq \frac{(K+1)(K-1)}{Kt^2} \leq \frac{K}{t^2}.
$$

□

D.2 Lower Bounds

Proof of Theorem 2. In the following, we will establish the asymptotic and minimax lower bounds, respectively.

Asymptotic lower bound. Consider any algorithm π that is R^{RG}-consistent. Sine $\mathbb{E}[N_i^{(0)}(T)] + \mathbb{E}[N_i^{(1)}(T)] = \mathbb{E}[N_i(T)]$ for all arms $i \in [K]$, we can utilize the regret decomposition in (4) to derive

$$
\begin{aligned}
R_{\mu,c}^{\mathrm{RG}}(T,\pi) &\geq \sum_{i>1}\left(\Delta_i \cdot \mathbb{E}[N_i^{(0)}(T)] + c\cdot \mathbb{E}[N_i^{(1)}(T)]\right) \\
&\geq \sum_{i>1}\left((\Delta_i \wedge c)\cdot \mathbb{E}[N_i(T)]\right).
\end{aligned}
$$

Fix any abstention regret $c > 0$. Then for all bandit instances $\mu \in \mathcal{U}$ and $a > 0$, it holds that

$$
\begin{aligned}
R_{\mu}^{\mathrm{CA}}(T,\pi) &= T\mu_1 - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] \\
&= \sum_{i>1}\left(\Delta_i \cdot \mathbb{E}[N_i(T)]\right) \\
&\leq \max_{i>1}\frac{\Delta_i}{\Delta_i \wedge c}\cdot \sum_{i>1}\left((\Delta_i \wedge c)\cdot \mathbb{E}[N_i(T)]\right) \\
&\leq \max_{i>1}\frac{\Delta_i}{\Delta_i \wedge c}\cdot R_{\mu,c}^{\mathrm{RG}}(T,\pi) \\
&= o(T^a).
\end{aligned}
$$

Therefore, in accordance with Definition 3, the algorithm π is also R^{CA}-consistent for arbitrary abstention regret c.

Subsequently, for any abstention regret c and bandit instance μ, we have

$$\begin{aligned}\liminf_{T\to\infty} \frac{R^{\mathrm{RG}}_{\mu,c}(T,\pi)}{\log T} &\geq \liminf_{T\to\infty} \frac{\sum_{i>1}\left((\Delta_i \wedge c)\cdot \mathbb{E}[N_i(T)]\right)}{\log T} \\ &= \sum_{i>1}(\Delta_i \wedge c)\cdot \liminf_{T\to\infty} \frac{\mathbb{E}[N_i(T)]}{\log T} \\ &\geq 2\sum_{i>1} \frac{\Delta_i \wedge c}{\Delta_i^2},\end{aligned}$$

where the last inequality follows from the property of R^{CA}-consistent policies as detailed in Remark 8.

This concludes the proof of the instance-dependent asymptotic lower bound.

Minimax lower bound. We extend the proof of the minimax lower bound from the canonical multi-armed bandit model to the model incorporating fixed-regret abstention.

Consider any fixed abstention regret $c > 0$, time horizon $T \geq K$ and algorithm $\pi \in \Pi^{\mathrm{RG}}$. We construct a bandit instance $\mu \in \mathcal{U}$, where $\mu_1 = \Delta$ and $\mu_i = 0$ for all $i \in [K] \setminus \{1\}$. Here, $\Delta > 0$ is some parameter whose exact value will be determined later. We use $\mathbb{P}_{\mu,c}$ to represent the probability distribution of the sequence $(A_1, B_1, X_1, \ldots, A_T, B_T, X_T)$ induced by the algorithm π for the abstention regret c and bandit instance μ. Since $\sum_{i=1} \mathbb{E}_{\mu,c}[N_i(T)] = T$, according to the pigeonhole principle, there must exist an index $j \in [K] \setminus \{1\}$ such that

$$\mathbb{E}_{\mu,c}[N_j(T)] \leq \frac{T}{K-1}.$$

Now we construct another bandit instance $\mu' \in \mathcal{U}$, where $\mu'_1 = \Delta$, $\mu'_j = 2\Delta$ and $\mu'_i = 0$ for all $i \in [K] \setminus \{1, j\}$. Let $\mathbb{P}_{\mu',c}$ denote the probability distribution of the sequence $(A_1, B_1, X_1, \ldots, A_T, B_T, X_T)$ induced by the algorithm π for the abstention regret c and bandit instance μ'.

For the first bandit instance μ, regardless of the abstention option, if $N_1(T) \leq T/2$, then the cumulative regret must be at least $(\Delta \wedge c)T/2$. Therefore, we have

$$R^{\mathrm{RG}}_{\mu,c}(T,\pi) \geq \frac{(\Delta \wedge c)T}{2} \mathbb{P}_{\mu,c}(N_1(T) \leq T/2).$$

Similarly, for the second bandit instance μ', we can obtain

$$R^{\mathrm{RG}}_{\mu',c}(T,\pi) \geq \frac{(\Delta \wedge c)T}{2} \mathbb{P}_{\mu',c}(N_1(T) > T/2).$$

By combining the aforementioned two inequalities and applying Lemma 1, we obtain the following:

$$\begin{aligned}R^{\mathrm{RG}}_{\mu,c}(T,\pi) + R^{\mathrm{RG}}_{\mu',c}(T,\pi) &\geq \frac{(\Delta \wedge c)T}{2}\left(\mathbb{P}_{\mu,c}(N_1(T) \leq T/2) + \mathbb{P}_{\mu',c}(N_1(T) > T/2)\right) \\ &\geq \frac{(\Delta \wedge c)T}{4} \exp\left(-\mathrm{KL}\left(\mathbb{P}_{\mu,c}, \mathbb{P}_{\mu',c}\right)\right).\end{aligned}$$

Leveraging Lemma 2 and the KL divergence between Gaussian distributions, we can derive

$$\mathrm{KL}\left(\mathbb{P}_{\mu,c}, \mathbb{P}_{\mu',c}\right) = \mathbb{E}_{\mu,c}[N_j(T)] \frac{(2\Delta)^2}{2} \leq \frac{2T\Delta^2}{K-1}.$$

Altogether, we can arrive at

$$R^{\mathrm{RG}}_{\mu,c}(T,\pi) + R^{\mathrm{RG}}_{\mu',c}(T,\pi) \geq \frac{(\Delta \wedge c)T}{4} \exp\left(-\frac{2T\Delta^2}{K-1}\right).$$

Now, we set $\Delta = \sqrt{\frac{K}{T}} \wedge c$, which leads to

$$\begin{aligned} R^{\mathrm{RG}}_{\mu,c}(T,\pi) + R^{\mathrm{RG}}_{\mu',c}(T,\pi) &\geq \frac{\sqrt{KT} \wedge cT}{4} \exp\left(-\frac{2T(\sqrt{\frac{K}{T}} \wedge c)^2}{K-1}\right) \\ &\geq \frac{\sqrt{KT} \wedge cT}{4} \exp\left(-\frac{2K}{K-1}\right) \\ &\geq \frac{\exp(-4)}{4}(\sqrt{KT} \wedge cT). \end{aligned}$$

Consequently, either $R^{\mathrm{RG}}_{\mu,c}(T,\pi)$ or $R^{\mathrm{RG}}_{\mu',c}(T,\pi)$ is at least $\frac{\exp(-4)}{8}(\sqrt{KT} \wedge cT)$, which completes the proof of the instance-independent minimax lower bound. □

D.3 Heterogeneous Abstention Regret

In the scenario of heterogeneous abstention regret, for each arm $i \in [K]$, the agent incurs a regret of $c_i > 0$ when pulling arm i and choosing abstention. By the law of total expectation, the expected cumulative regret up to the time horizon T can be expressed as:

$$\begin{aligned} R^{\mathrm{RG}}_{\mu,\mathbf{c}}(T,\pi) &= \mathbb{E}\left[\sum_{t=1}^{T} \left((\mu_1 - X_t) \cdot \mathbf{1}\{B_t = 0\} + c_{A_t} \cdot \mathbf{1}\{B_t = 1\}\right)\right] \\ &= c_1 \cdot \mathbb{E}[N_1^{(1)}(T)] + \sum_{i>1} \left(\Delta_i \cdot \mathbb{E}[N_i^{(0)}(T)] + c_i \cdot \mathbb{E}[N_i^{(1)}(T)]\right). \end{aligned}$$

For ease of presentation, we define $c = \max_{i\in[K]} c_i$ as the maximum abstention regret level. The lower bound part is straightforward. Specifically, we can obtain the following lower bounds.

Theorem 6. *For any abstention regret $\mathbf{c} = (c_i)_{i\in[K]}$, bandit instance $\mu \in \mathcal{U}$ and R^{RG}-consistent algorithm π, it holds that*

$$\liminf_{T\to\infty} \frac{R^{\mathrm{RG}}_{\mu,\mathbf{c}}(T,\pi)}{\log T} \geq 2\sum_{i>1} \frac{\Delta_i \wedge c_i}{\Delta_i^2}.$$

For any maximum abstention regret level $c > 0$ and time horizon $T \geq K$, there exists a universal constant $\alpha > 0$ such that

$$\inf_{\pi\in\Pi^{\mathrm{RG}}} \sup_{\mu\in\mathcal{U},\mathbf{c}} R^{\mathrm{RG}}_{\mu,\mathbf{c}}(T,\pi) \geq \alpha(\sqrt{KT} \wedge cT).$$

With respect to the algorithm design, for each arm $i \in [K]$, we redefine its lower confidence bound as

$$\mathrm{LCB}_i(t) = \hat{\mu}_i(t-1) - \sqrt{\frac{6\log t + 2\log(c_i \vee 1)}{N_i(t-1)}}.$$

Then in the abstention decision rule, we choose to abstain ($B_t = 1$) if and only if

$$\max_{i\in[K]\setminus\{A_t\}} \mathrm{LCB}_i(t) - \hat{\mu}_{A_t}(t-1) \geq c_i \quad \text{or} \quad \sqrt{\frac{K}{t}} \geq c. \tag{18}$$

Note that in (18), we compare $\sqrt{\frac{K}{t}}$ with $c = \max_{i\in[K]} c_i$.

Theorem 7. *For any abstention regret $\mathbf{c} = (c_i)_{i\in[K]}$ and bandit instance $\mu \in \mathcal{U}$, Algorithm 1 with the new abstention criteria in* (18) *guarantees that*

$$\limsup_{T\to\infty} \frac{R^{\mathrm{RG}}_{\mu,\mathbf{c}}(T)}{\log T} \leq 2\sum_{i>1} \frac{\Delta_i \wedge c_i}{\Delta_i^2}.$$

Furthermore, there exists a universal constant $\alpha > 0$ such that

$$R^{\mathrm{RG}}_{\mu,\mathbf{c}}(T) \leq \begin{cases} cT & \text{if } c \leq \sqrt{K/T} \\ \alpha(\sqrt{KT} + \sum_{i>1} \Delta_i) & \text{if } c > \sqrt{K/T}. \end{cases}$$

The proof of Theorem 7 closely follows that of Theorem 1 in Appendix D.1. While many steps are analogous, it is noteworthy that in this context, for any arm i with $\Delta_i < c_i$, we can upper bound $\mathbb{E}[N_i^{(1)}(T)]$ as

$$\begin{aligned}
\mathbb{E}[N_i^{(1)}(T)] \leq\ & \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left\{A_t = i \text{ and } \sqrt{\frac{K}{t}} \geq c\right\}\right] \\
& + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{A_t = i \text{ and } \max_{j\in[K]}\left(\hat{\mu}_j(t-1) - \sqrt{\frac{6\log t + 2\log(c_j \vee 1)}{N_j(t-1)}}\right) \geq \mu_1\right\}\right] \\
& + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{I}\left\{A_t = i \text{ and } \mu_1 - \hat{\mu}_i(t-1) \geq c_i\right\}\right].
\end{aligned}$$

E Analysis of the Fixed-Reward Setting

E.1 Upper Bounds

Proof of Theorem 3. Utilizing the law of total expectation, we can decompose the regret $R^{\mathrm{RW}}_{\mu,c}(T,\pi)$ in the following:

$$\begin{aligned}
R^{\mathrm{RW}}_{\mu,c}(T,\pi) &= T\cdot(\mu_1 \vee c) - \mathbb{E}\left[\sum_{t=1}^{T}\Big(X_t \cdot \mathbf{1}\{B_t = 0\} + c\cdot \mathbf{1}\{B_t = 1\}\Big)\right] \\
&= T\cdot(\mu_1 \vee c) - \mathbb{E}\left[\sum_{t=1}^{T}\Big(\mu_{A_t} \cdot \mathbf{1}\{B_t = 0\} + c\cdot \mathbf{1}\{B_t = 1\}\Big)\right] \\
&= \sum_{i\in[K]}\Big((\mu_1 \vee c - \mu_i)\cdot \mathbb{E}[N_i^{(0)}(T)] + (\mu_1 \vee c - c)\cdot \mathbb{E}[N_i^{(1)}(T)]\Big). \qquad (19)
\end{aligned}$$

Recall that we define $\hat{\mu}_i(t) = +\infty$ if $N_i(t) = 0$ for all arms $i \in [K]$. Thus, for any arm i with $\mu_i > c$, we can obtain

$$\begin{aligned}
\mathbb{E}[N_i^{(1)}(T)] &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{A_t = i \text{ and } B_t = 1\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{A_t = i \text{ and } \hat{\mu}_i(t-1) \leq c\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=0}^{T-1}\mathbb{I}\{A_t = i \text{ and } \hat{\mu}_{is} \leq c \text{ and } N_i(t-1) = s\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=1}^{T-1}\mathbb{I}\{A_t = i \text{ and } \hat{\mu}_{is} \leq c \text{ and } N_i(t-1) = s\}\right] \\
&\leq \mathbb{E}\left[\sum_{s=1}^{T-1}\mathbb{I}\{\hat{\mu}_{is} \leq c\}\right] \\
&\leq \frac{2}{(\mu_i - c)^2}
\end{aligned}$$

where the penultimate inequality is derived from an argument analogous to that in (6), and the last inequality is a consequence of Lemma 4.

Similarly, for any arm i with $\mu_i < c$, we have

$$\begin{aligned}
\mathbb{E}[N_i^{(0)}(T)] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{A_t = i \text{ and } B_t = 0\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{A_t = i \text{ and } \hat{\mu}_i(t-1) > c\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=0}^{T-1} \mathbb{I}\{A_t = i \text{ and } \hat{\mu}_{is} > c \text{ and } N_i(t-1) = s\}\right] \\
&\leq 1 + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=1}^{T-1} \mathbb{I}\{A_t = i \text{ and } \hat{\mu}_{is} > c \text{ and } N_i(t-1) = s\}\right] \\
&\leq 1 + \frac{2}{(c-\mu_i)^2}.
\end{aligned}$$

Asymptotic upper bound. First, we consider the scenario where $\mu_1 \leq c$. In this case, based on the regret decomposition in Equation (19), we can bound the regret as follows:

$$\begin{aligned}
R_{\mu,c}^{\mathrm{RW}}(T) &= \sum_{i\in[K]} (c-\mu_i)\cdot \mathbb{E}[N_i^{(0)}(T)] \\
&= \sum_{i:\mu_i<c} (c-\mu_i)\cdot \mathbb{E}[N_i^{(0)}(T)] \\
&\leq \sum_{i:\mu_i<c} \left(c-\mu_i+\frac{2}{c-\mu_i}\right) \\
&= o(\log T).
\end{aligned} \tag{20}$$

Next, we consider the scenario where $\mu_1 > c$. Due to the asymptotic optimality of the base algorithm, for any suboptimal arm i, we have

$$\mathbb{E}[N_i(T)] \leq \frac{2\log T}{\Delta_i^2} + o(\log T).$$

Thus, we can bound the regret as:

$$\begin{aligned}
&R_{\mu,c}^{\mathrm{RW}}(T) \\
&= \sum_{i\in[K]} \Big((\mu_1-\mu_i)\cdot \mathbb{E}[N_i^{(0)}(T)] + (\mu_1 - c)\cdot \mathbb{E}[N_i^{(1)}(T)]\Big) \\
&= \sum_{i\in[K]} \Big((\mu_1-\mu_i\vee c+\mu_i\vee c-\mu_i)\cdot \mathbb{E}[N_i^{(0)}(T)] + (\mu_1-\mu_i\vee c+\mu_i\vee c-c)\cdot \mathbb{E}[N_i^{(1)}(T)]\Big) \\
&= \sum_{i\in[K]} \Big((\mu_1-\mu_i\vee c)\cdot \mathbb{E}[N_i(T)] + (\mu_i\vee c-\mu_i)\cdot \mathbb{E}[N_i^{(0)}(T)] + (\mu_i\vee c-c)\cdot \mathbb{E}[N_i^{(1)}(T)]\Big) \\
&= \sum_{i>1}(\mu_1-\mu_i\vee c)\cdot \mathbb{E}[N_i(T)] + \sum_{i:\mu_i<c}(c-\mu_i)\cdot \mathbb{E}[N_i^{(0}(T)] + \sum_{i:\mu_i>c}(\mu_i-c)\cdot \mathbb{E}[N_i^{(1)}(T)] \\
&\leq (2\log T)\sum_{i>1}\frac{\mu_1-\mu_i\vee c}{\Delta_i^2} + \sum_{i:\mu_i<c}\left(c-\mu_i+\frac{2}{c-\mu_i}\right) + \sum_{i:\mu_i>c}\frac{2}{\mu_i-c} + o(\log T) \\
&= (2\log T)\sum_{i>1}\frac{\mu_1-\mu_i\vee c}{\Delta_i^2} + o(\log T)
\end{aligned}$$

where the third equality is due to the fact that $\mathbb{E}[N_i^{(0)}(T)] + \mathbb{E}[N_i^{(1)}(T)] = \mathbb{E}[N_i(T)]$ for all arms $i \in [K]$.

Altogether, in both scenarios, it holds that

$$\limsup_{T\to\infty} \frac{R^{\mathrm{RW}}_{\mu,c}(T)}{\log T} \le 2\sum_{i>1} \frac{\mu_1 \vee c - \mu_i \vee c}{\Delta_i^2}.$$

Minimax upper bound. First, if $\mu_1 \le c$, by utilizing the regret decomposition in Equation (20), we have

$$\begin{aligned}
R^{\mathrm{RW}}_{\mu,c}(T) &= \sum_{i:\mu_i<c} (c-\mu_i)\cdot \mathbb{E}[N_i^{(0)}(T)] \\
&= \sum_{i:0<c-\mu_i<\sqrt{\frac{K}{T}}} (c-\mu_i)\cdot \mathbb{E}[N_i^{(0)}(T)] + \sum_{i:c-\mu_i\ge\sqrt{\frac{K}{T}}} (c-\mu_i)\cdot \mathbb{E}[N_i^{(0)}(T)] \\
&\le \sqrt{\frac{K}{T}} \sum_{i:0<c-\mu_i<\sqrt{\frac{K}{T}}} \mathbb{E}[N_i^{(0)}(T)] + \sum_{i:c-\mu_i\ge\sqrt{\frac{K}{T}}} \left(c-\mu_i+\frac{2}{c-\mu_i}\right) \\
&\le \sqrt{\frac{K}{T}}\cdot T + \sum_{i:c-\mu_i\ge\sqrt{\frac{K}{T}}} \left(c-\mu_i+2\sqrt{\frac{T}{K}}\right) \\
&\le \sqrt{KT} + \sum_{i\in[K]} (c-\mu_i) + K\cdot 2\sqrt{\frac{T}{K}} \\
&= 3\sqrt{KT} + \sum_{i\in[K]} (\mu_1 \vee c - \mu_i).
\end{aligned} \tag{21}$$

Next, if $\mu_1 > c$, then the best possible (expected) reward at a single time step is μ_1, which coincides with the canonical multi-armed bandit problem. Consequently, compared with canonical multi-armed bandits, at a single time step, the agent in our problem incurs a greater (expected) regret only if an arm i with $\mu_i > c$ is pulled and the abstention option is chosen. Thus, we have

$$R^{\mathrm{RW}}_{\mu,c}(T) \le R^{\mathrm{CA}}_{\mu}(T) + \sum_{i:\mu_i>c} (\mu_i - c)\cdot \mathbb{E}[N_i^{(1)}(T)].$$

Due to the minimax optimality of the base algorithm, there exists a universal constant $\alpha_1 > 0$ such that

$$R^{\mathrm{CA}}_{\mu}(T) \le \alpha_1 \left(\sqrt{KT} + \sum_{i>1} \Delta_i\right).$$

Furthermore, using a similar argument as in (21), we can derive

$$\begin{aligned}
&\sum_{i:\mu_i>c} (\mu_i - c)\cdot \mathbb{E}[N_i^{(1)}(T)] \\
&= \sum_{i:0<\mu_i-c<\sqrt{\frac{K}{T}}} (\mu_i-c)\cdot \mathbb{E}[N_i^{(1)}(T)] + \sum_{i:\mu_i-c\ge\sqrt{\frac{K}{T}}} (\mu_i-c)\cdot \mathbb{E}[N_i^{(1)}(T)] \\
&\le \sqrt{KT} + \sum_{i:\mu_i-c\ge\sqrt{\frac{K}{T}}} \frac{2}{\mu_i - c} \\
&\le 3\sqrt{KT}.
\end{aligned}$$

Therefore, we can bound the regret as

$$\begin{aligned} R^{\mathrm{RW}}_{\mu,c}(T) &\le (\alpha_1+3)\left(\sqrt{KT}+\sum_{i>1}\Delta_i\right) \\ &= (\alpha_1+3)\left(\sqrt{KT}+\sum_{i\in[K]}(\mu_1\vee c-\mu_i)\right). \end{aligned}$$

As a result, the desired minimax upper bound holds in both scenarios. □

E.2 Lower Bounds

Proof of Theorem 4. The proof structure for Theorem 4 closely parallels that of Theorem 2 in Appendix D.2, although certain specific details contain significant variations. Therefore, we will streamline the shared components and elaborate on the distinctions.

Asymptotic lower bound. Consider any R^{RW}-consistent algorithm π and bandit instance $\mu \in \mathcal{U}$. The case that $c \ge \mu_1$ is trivial, as $R^{\mathrm{RW}}_{\mu,c}(T,\pi)$ is non-negative, and $\mu_1 \vee c - \mu_i \vee c = 0$ for all $i > 1$. Thus, it suffices to demonstrate that for any abstention reward $c < \mu_1$,

$$\liminf_{T\to\infty}\frac{R^{\mathrm{RW}}_{\mu,c}(T,\pi)}{\log T} \ge 2\sum_{i>1}\frac{\mu_1-\mu_i\vee c}{\Delta_i^2}.$$

When $c < \mu_1$, we can establish a lower bound on $R^{\mathrm{RW}}_{\mu,c}(T,\pi)$ as follows:

$$\begin{aligned} R^{\mathrm{RW}}_{\mu,c}(T,\pi) &= \sum_{i\in[K]}\Big((\mu_1-\mu_i)\cdot\mathbb{E}[N_i^{(0)}(T)]+(\mu_1-c)\cdot\mathbb{E}[N_i^{(1)}(T)]\Big) \\ &\ge \sum_{i>1}\Big((\mu_1-\mu_i)\cdot\mathbb{E}[N_i^{(0)}(T)]+(\mu_1-c)\cdot\mathbb{E}[N_i^{(1)}(T)]\Big) \\ &\ge \sum_{i>1}\Big((\mu_1-\mu_i\vee c)\cdot\mathbb{E}[N_i(T)]\Big). \end{aligned}$$

Therefore, we only need to show for all suboptimal arms $i > 1$,

$$\liminf_{T\to\infty}\frac{\mathbb{E}[N_i(T)]}{\log T} \ge \frac{2}{\Delta_i^2}. \tag{22}$$

However, unlike the asymptotic lower bound part of the proof of Theorem 2, we cannot apply the properties of R^{CA}-consistency here, as R^{RW}-consistency does not imply R^{CA}-consistency in general. Instead, we will demonstrate the desired result (22) directly.

Fix an index $j > 1$ and take $\varepsilon > 0$. We now proceed to create an alternative bandit instance $\mu' \in \mathcal{U}$, where $\mu'_j = \mu_1 + \varepsilon$ and $\mu'_i = \mu_i$ for all $i \in [K] \setminus \{j\}$. Note that for the new bandit instance, it holds that $\max_{i\in[K]}\mu'_i = \mu'_j > \mu_1 > c$. To distinguish between the two scenarios, we will refer to the probability distribution associated with the sequence $(A_1, B_1, X_1, \ldots, A_T, B_T, X_T)$, generated by the algorithm π for the abstention reward c and the original bandit scenario μ, as $\mathbb{P}_{\mu,c}$, and for the new bandit instance μ', we denote the corresponding distribution as $\mathbb{P}_{\mu',c}$.

A straightforward computation yields

$$R^{\mathrm{RW}}_{\mu,c}(T,\pi) \ge \frac{(\mu_1-\mu_i\vee c)T}{2}\mathbb{P}_{\mu,c}(N_j(T) > T/2)$$

and

$$R^{\mathrm{RW}}_{\mu',c}(T,\pi) \ge \frac{\varepsilon T}{2}\mathbb{P}_{\mu',c}(N_j(T) \le T/2).$$

Employing a similar approach to the one used in the minimax lower bound part of the proof of Theorem 2, utilizing Lemmas 1 and 2, we can derive:

$$\begin{aligned} R^{\mathrm{RW}}_{\mu,c}(T,\pi) + R^{\mathrm{RW}}_{\mu',c}(T,\pi) &\geq \frac{((\mu_1 - \mu_i \vee c) \wedge \varepsilon)T}{2} \left(\mathbb{P}_{\mu,c}(N_j(T) > T/2) + \mathbb{P}_{\mu',c}(N_j(T) \leq T/2)\right) \\ &\geq \frac{((\mu_1 - \mu_i \vee c) \wedge \varepsilon)T}{4} \exp\left(-\mathrm{KL}\left(\mathbb{P}_{\mu,c}, \mathbb{P}_{\mu',c}\right)\right) \\ &= \frac{((\mu_1 - \mu_i \vee c) \wedge \varepsilon)T}{4} \exp\left(-\mathbb{E}_{\mu,c}[N_j(T)] \frac{(\Delta_j + \varepsilon)^2}{2}\right). \end{aligned}$$

By rearranging the above inequality and taking the limit inferior, we have

$$\begin{aligned} \liminf_{T\to\infty} \frac{\mathbb{E}[N_j(T)]}{\log T} &\geq \frac{2}{(\Delta_j + \varepsilon)^2} \left(1 + \limsup_{T\to\infty} \frac{\log\left(\frac{(\mu_1 - \mu_i \vee c) \wedge \varepsilon}{4\left(R^{\mathrm{RW}}_{\mu,c}(T,\pi) + R^{\mathrm{RW}}_{\mu',c}(T,\pi)\right)}\right)}{\log T}\right) \\ &= \frac{2}{(\Delta_j + \varepsilon)^2} \left(1 - \limsup_{T\to\infty} \frac{\log\left(R^{\mathrm{RW}}_{\mu,c}(T,\pi) + R^{\mathrm{RW}}_{\mu',c}(T,\pi)\right)}{\log T}\right). \end{aligned}$$

Recall the definition of R^{RW}-consistency. For all $a > 0$, both $R^{\mathrm{RW}}_{\mu,c}(T,\pi)$ and $R^{\mathrm{RW}}_{\mu',c}(T,\pi)$ are on the order of $o(T^a)$, and hence,

$$\limsup_{T\to\infty} \frac{\log\left(R^{\mathrm{RW}}_{\mu,c}(T,\pi) + R^{\mathrm{RW}}_{\mu',c}(T,\pi)\right)}{\log T} \leq a.$$

By letting both a and ε approach zero, we can establish the desired result (22), thereby concluding the proof of the asymptotic lower bound.

Minimax lower bound. The construction employed in the fixed-reward setting here is analogous to the one utilized in the proof of Theorem 2.

Consider any fixed abstention reward $c \in \mathbb{R}$, time horizon $T \geq K$ and algorithm $\pi \in \Pi^{\mathrm{RW}}$. Let $\Delta > 0$ be a parameter to be determined later. We construct a bandit instance $\mu \in \mathcal{U}$, where $\mu_1 = \Delta + c$ and $\mu_i = c$ for all $i \in [K] \setminus \{1\}$. Note that there must exist an index $j \in [K] \setminus \{1\}$ such that $\mathbb{E}_{\mu,c}[N_j(T)] \leq \frac{T}{K-1}$. We then construct another bandit instance $\mu' \in \mathcal{U}$, where $\mu'_1 = \Delta + c$, $\mu'_j = 2\Delta + c$ and $\mu'_i = c$ for all $i \in [K] \setminus \{1, j\}$.

Similarly, by applying Lemmas 1 and 2, we can derive that

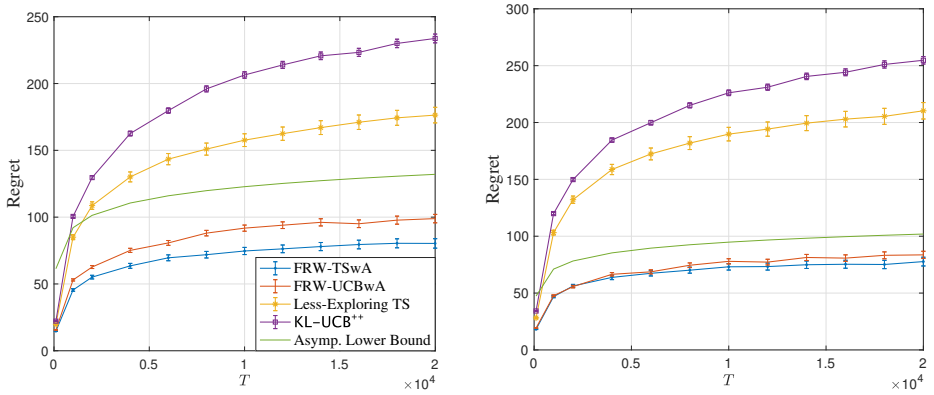
$$\begin{aligned} R^{\mathrm{RW}}_{\mu,c}(T,\pi) + R^{\mathrm{RW}}_{\mu',c}(T,\pi) &\geq \frac{\Delta T}{2} \left(\mathbb{P}_{\mu,c}(N_1(T) \leq T/2) + \mathbb{P}_{\mu',c}(N_1(T) > T/2)\right) \\ &\geq \frac{\Delta T}{4} \exp\left(-\mathrm{KL}\left(\mathbb{P}_{\mu,c}, \mathbb{P}_{\mu',c}\right)\right) \\ &= \frac{\Delta T}{4} \exp\left(-\mathbb{E}_{\mu,c}[N_j(T)] \frac{(2\Delta)^2}{2}\right) \\ &\geq \frac{\Delta T}{4} \exp\left(-\frac{2T\Delta^2}{K-1}\right). \end{aligned}$$

By choosing $\Delta = \sqrt{\frac{K}{T}}$, we have

$$R^{\mathrm{RW}}_{\mu,c}(T,\pi) + R^{\mathrm{RW}}_{\mu',c}(T,\pi) \geq \frac{\exp(-4)}{8}\sqrt{KT}.$$

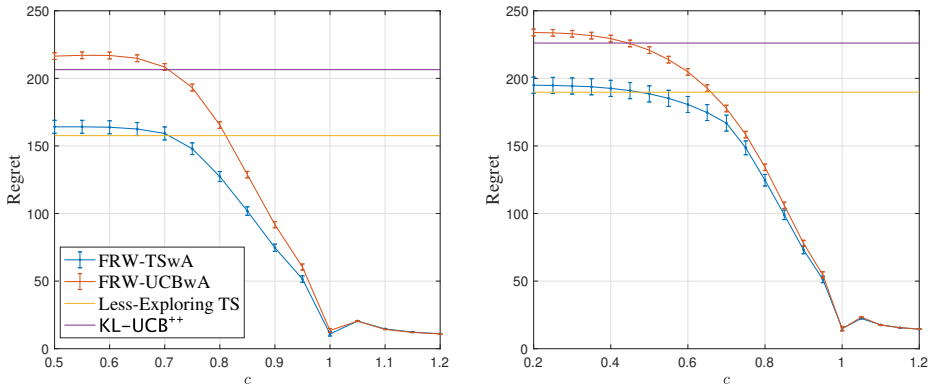
Consequently, either $R^{\mathrm{RW}}_{\mu,c}(T,\pi)$ or $R^{\mathrm{RW}}_{\mu',c}(T,\pi)$ is at least $\frac{\exp(-4)}{8}\sqrt{KT}$.

Therefore, we have established the instance-independent minimax lower bound. □



(a) Instance $\mu^\dagger$ (b) Instance $\mu^\ddagger$

Figure 4: Empirical regrets with abstention reward $c = 0.9$ for different time horizons T.



(a) Instance $\mu^\dagger$ (b) Instance $\mu^\ddagger$

Figure 5: Empirical regrets with time horizon $T = 10,000$ for different abstention rewards c.

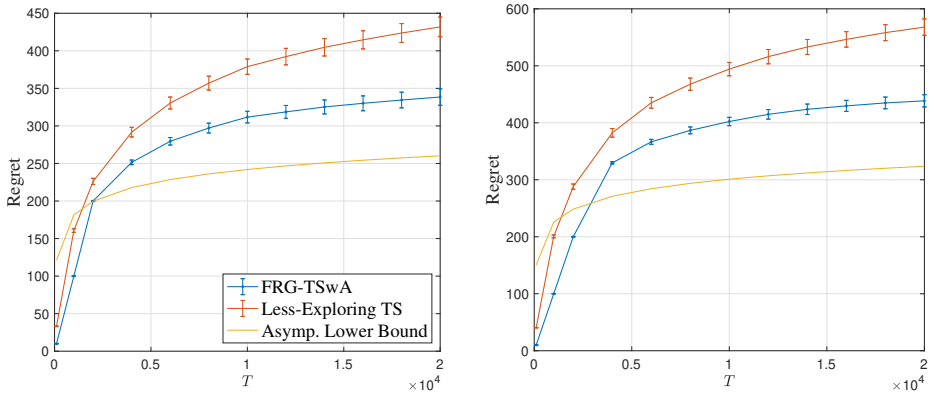
F Numerical Experiments

F.1 Results for the Fixed-Reward Setting

In this subappendix, we present the empirical results pertaining to the fixed-reward setting. Specifically, we examine the empirical performances of two particular realizations of our algorithm FRW-ALGwA (as outlined in Algorithm 2): FRW-TSwA and FRW-UCBwA. The former uses Less-Exploring Thompson Sampling (Jin et al., 2023) as its base algorithm, while the latter employs KL-UCB^{++} (Ménard & Garivier, 2017). Note that KL-UCB^{++} is not an anytime algorithm; that is, it requires the prior knowledge of the time horizon T as an input parameter. In the following experiments, we continue to utilize the two bandit instances $\mu^\dagger$ and $\mu^\ddagger$, as previously defined in Appendix F.1.

In a manner analogous to our methodology for the fixed-regret setting, we adopt the original versions of Less-Exploring TS and KL-UCB^{++} as baseline algorithms without the abstention option. The experimental results of the different methods with abstention reward $c = 0.9$ for different time horizons T are presented in Figure 4. Additionally, we plot the instance-dependent asymptotic lower bound (ignoring the limit in T) on the cumulative regret (see Theorem 4) within each sub-figure. From Figure 4, we have the following observations:

- Both realizations of our algorithm, FRW-TSwA and FRW-UCBwA, exhibit marked superiority over the two non-abstaining baselines.



(a) $K = 20$ (b) $K = 30$

Figure 6: Empirical regrets with abstention regret $c = 0.1$ for different time horizons T.

- Concerning the observed growth trend, as the time horizon T increases, the performance curves for both FRW-TSwA and FRW-UCBwA approximate the asymptotic lower bound closely. This behavior indicates that the expected cumulative regrets of FRW-TSwA and FRW-UCBwA attain the instance-dependent lower bound asymptotically, validating the theoretical findings discussed in Section 4.
- While both FRW-TSwA and FRW-UCBwA represent implementations of our general algorithm and share identical theoretical guarantees, FRW-TSwA demonstrates superior empirical performance. This is particularly evident in the first instance $\mu^\dagger$, suggesting its enhanced applicability for real-world applications.

Next, we examine the impact of the abstention reward c by assessing the performance of FRW-TSwA and FRW-UCBwA for different c, while keeping the time horizon T fixed at $10,000$. The experimental results for bandit instances $\mu^\dagger$ and $\mu^\ddagger$ are shown in Figure 5.

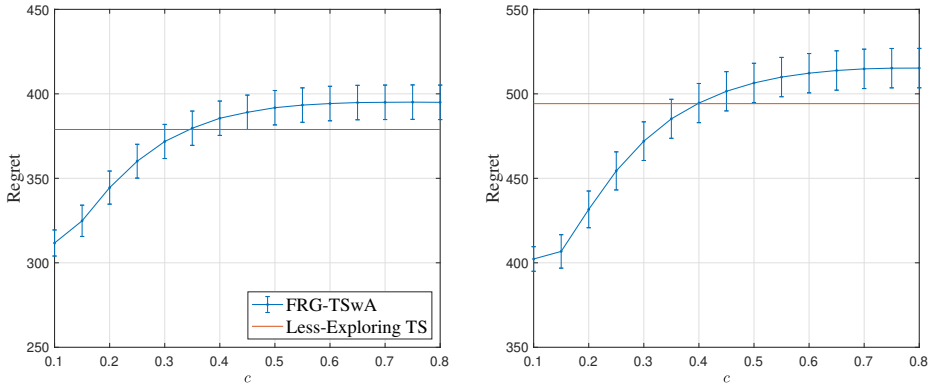
Within each sub-figure, a pattern emerges. As the abstention reward c increases, the empirical average cumulative regret initially remains relatively stable and starts to decline once c crosses a certain threshold, eventually stabilizing around a small value. These observations are consistent with our theoretical expectations. Specifically, when the abstention reward c is lower than the smallest mean reward among the arms, the agent derives no benefit from opting for the abstention action over selecting an arm. On the other hand, when the abstention reward c exceeds the highest mean reward of the arms, abstention becomes the optimal decision and its reward is even superior to choosing the best arm. In this specific scenario, it is possible to achieve a regret of $o(\log T)$; see Remark 6 for further insights.

F.2 Random Instances

In this subappendix, we present additional numerical experiments, using random instances with large numbers of arms. The construction of these random instances mirrors the method in Jin et al. (2023). Specifically, for a given number of arms denoted by $K \geq 10$, we set $\mu_1 = 1$ and $\mu_i = 0.7$ for $i \in \{2, 3, \ldots, 10\}$, while $\mu_i \sim \text{Unif}[0.3, 0.5]$ for $i \in [K] \setminus [10]$.

For the sake of simplicity in presentation, we focus on the fixed-regret setting, examining two choices of K, namely, $K = 20$ and $K = 30$. The empirical averaged cumulative regrets with abstention regret $c = 0.1$ for different time horizons T are shown in Figure 6, while the experimental results with time horizon $T = 10,000$ for different abstention rewards c are illustrated in Figure 7.

It is evident that the findings in Figures 6 and 7 closely resemble those in Figures 2 and 3. Notably, FRW-TSwA outperforms Less-Exploring TS which is not tailored to the setting with the abstention option.



(a) $K = 20$ (b) $K = 30$

Figure 7: Empirical regrets with time horizon $T = 10,000$ for different abstention regrets c.