On Multi-Armed Bandit with Impatient Arms

Yuming Shao [1 2] **Zhixuan Fang** [1 2]

Abstract

In this paper, we investigate a Multi-Armed Bandit (MAB) setting where an arm exits the game if the algorithm continuously neglects it. This setup is motivated by real-world scenarios, such as online advertising and crowdsourcing, where arms only gain benefits after being pulled by the algorithm. We identify the intrinsic hardness of this problem and limitations in existing approaches. We propose FC-SE algorithm with expected regret upper bounds as our solution to this problem. As an extension, we even allow new arms to enter after the game starts and design FC-Entry algorithm with performance guarantees for this setup. Finally, we conduct experiments to validate our theoretical results.

1. Introduction

Multi-Armed Bandit (MAB), first introduced in (Robbins, 1952), is a type of machine learning model used to describe the decision-making problems with unknown information that needs to be learned. In this model, each arm represents a potential choice with unknown expected reward. The algorithm's objective is to achieve the highest possible overall reward. The central challenge in a MAB problem is to strike a balance between exploiting the arms that have yielded high rewards and exploring unknown arms to uncover their potential. MAB algorithms find real-world applications in various fields, such as online advertising (Schwartz et al., 2017; Yang & Lu, 2016; Aramayo et al., 2023; Avadhanula et al., 2021; Han & Gabor, 2020), clinical trials (Aziz et al., 2021; Chakravorty & Mahajan, 2014), recommendation systems (Santana et al., 2020; Xie et al., 2021; Mahadik et al., 2020), crowdsourcing (Rangi & Franceschetti, 2018; Song & Jin, 2021; Qin et al., 2023; Liu & Liu, 2017; Zhang et al., 2015; Tran-Thanh et al., 2014), and resource allocation in computer networks (Zuo & Joe-Wong, 2021; Pase et al., 2022; Feki & Capdevielle, 2011). In these applications, bandit algorithms attempt to find choices that maximize their own benefits, while arms passively comply with and accept the arrangements made by the algorithm.

However, in many situations, bandit arms also have their own interests, preferences, and even the right to make choices. For instance, a series of recent studies (Liu et al., 2020; 2021a; Kong et al., 2022; Basu et al., 2021; Sankararaman et al., 2021; Zhang et al., 2022) investigated multi-agent multi-armed bandit in two-sided matching markets. In these contexts, each arm selects its most preferred agent among those who have chosen it. Beyond this, the arms' preferences and choices can manifest in various ways. One crucial aspect is their decision to participate in the game. If participation promises satisfactory benefits, they may opt to remain involved; conversely, if participation does not yield sufficient benefits, they may be motivated to exit. To illustrate this, we present the following real world scenarios:

Example 1 (Online Advertising). A website owner allocates advertisement slots to advertisers across various time periods. The owner generally prefers advertisements with high click-through rates to generate higher revenue. However, for advertisers with low click-through rates, their ads may remain undisplayed for an extended period. This situation may lead to financial losses for advertisers and evoke their dissatisfaction. Consequently, advertisers may find it advantageous to withdraw from the competition for ad slots on this site.

Example 2 (Crowdsourcing). The crowdsourcing workers participating in a dataset labeling task receive compensation for each label they provide. A crowdsourcing system assuming control over the task assignments to workers (Zhang et al., 2015) often continuously assigns tasks to workers with proven proficiency, while potentially neglecting others. Given that workers are only compensated upon task completion, certain workers might experience prolonged periods without receiving any tasks, leading to a lack of income from the platform. They may lose patience and turn to alternative avenues for earning income.

What these two cases have in common is that if the algorithm frequently neglects an arm, it can harm the arm's interests, leading it to lose patience and ultimately incentivizing it to leave. Patience, to some extent, represents

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China [2] Shanghai Qi Zhi Institute, Shanghai, China. Correspondence to: Zhixuan Fang <zfang@mail.tsinghua.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

the threshold of loss that an arm is willing to tolerate. The conventional MAB model does not account for the arms' patience and the possibility that they may leave the game if they run out of patience, making it inadequate for effectively capturing the real-world examples outlined above. To address this limitation, we introduce the patience of arms into the conventional model and study the *Multi-Armed Bandit with Impatient Arms* setting. In this setup, there are K arms and each arm k is associated with a positive integer m_k indicating arm k's patience. If arm k is consistently ignored by the algorithm for m_k consecutive times, it will leave the game. Consequently, the algorithm cannot pull it within the remaining time horizon.

While existing bandit algorithms such as the Upper Confidence Bound (UCB) (Lai, 1987; Auer et al., 2002), Successive Elimination (SE) (Even-Dar et al., 2006) and Thompson Sampling (TS) (Thompson, 1933) algorithms can operate within the proposed setting, they do not explicitly accommodate the limited patience of the arms. When the optimal arm is impatient, it could deplete its patience and leave the game early. As a result, the algorithm would be compelled to choose the remaining sub-optimal arms, leading to a linear regret. Therefore, in this paper, we ask two research questions: *1) How does the patience of arms impact the performance of existing algorithms? 2) How can we address the challenges posed by the impatience of arms?*

1.1. Our Contributions

- We introduce a new version of MAB problem, where an arm leaves the game if the algorithm continuously neglects it. As an extension, we also consider a more general and realistic setup that allows not only the participating arms to exit but also new arms to enter.

- We derive a minimax lower bound to highlight the fundamental hardness of the proposed problem especially when the arms exhibit significant impatience. We also comprehensively study the existing algorithms such as UCB and SE in the proposed setup. We identify a broad range of problem instances where they incur (nearly) linear expected regret.

- When existing algorithms have no guaranteed performance, we propose the Feasible Cycle-based Successive Elimination (FC-SE) algorithm. In FC-SE, we repeat a special sequence of arms (referred to as a feasible cycle) to prevent unexpected arm departures. When it is possible to include all arms in the feasible cycle, FC-SE achieves a regret of $\tilde{O}\big(n + \sqrt{nT}\big)$, where n is the feasible cycle length and T is the time horizon. However, when it's only possible to include a subset of arms in the initial feasible cycle, FC-SE randomly drops arms from the current feasible cycle, if necessary, to accommodate remaining arms, achieving a regret of $\tilde{O}\big(K^{\frac{4}{3}}T^{\frac{2}{3}}n^{\frac{1}{3}}\big)$ if the remaining arms have sufficient patience. We design the FC-Entry algorithm for the extension scenario with new entering arms. FC-Entry accounts for unknown entry times for new arms, assuming that entries are sparse. FC-Entry achieves a dynamic regret of $\tilde{O}\big(K^{2}T^{\frac{2}{3}}n^{\frac{1}{3}}\big)$ if the arms initially available but not in the initial feasible cycle have sufficient patience.

- We conduct numerical experiments to validate our theoretical results.

1.2. Related Work

There are several prior works allowing time-variant arm sets. The literature on sleeping bandits (Kanade et al., 2009; Kanade & Steinke, 2014; Cortes et al., 2019; Saha et al., 2020; Gaillard et al., 2023) assumes that the set A_t of available arms at any given time t may change. Additionally, Chakrabarti et al. (2008) and Tracà et al. (2020) are related to our work since they assume that if an arm leaves the game, it will not return. While our setting is related to these works, there are significant differences. In these papers, arm availability can be either stochastic (Saha et al., 2020) or adversarial (Gaillard et al., 2023). On the one hand, our setting assumes that the available set of arms is determined by the algorithm's previous actions, making it incompatible with the stochastic arm availability assumption. On the other hand, in previous works assuming adversarial arm availability, the algorithm's performance is typically evaluated by comparing its choice at time t against an arm from the adversarially selected A_t. In our scenario, instead, the algorithm's choice should always be compared against the offline optimal arm k^*, even though an algorithm can unfortunately lose it ($k^* \notin A_t$).

One step in our proposed algorithms is to schedule the arms in a way that prevents any of them from running out of patience. Similar scheduling problems are considered in the Age of Information (AoI) literature (Kaul et al., 2011; 2012; Abbas et al., 2023). In the field of communication, AoI serves as a metric for measuring information freshness. Although a main line of research has focused on AoI minimization (Liu et al., 2019; Arafa et al., 2020; Chen et al., 2023), our paper is more related to scheduling under hard constraints. For instance, the peak AoI deadline resembles the concept of arm patience in this paper. Li et al. (2021); Liu et al. (2021b); Li et al. (2023a) study the scheduling problem under various AoI constraints. Due to space limits, we defer other details of related work to Appendix A.

2. Preliminaries

The problem studied in this paper is built upon the standard stochastic MAB. We consider a time horizon of length

T and a finite set of K arms. At each time $t \in [T]$, an algorithm pulls one arm $a_t \in A_t$, where A_t is the set of arms available at time t. When arm k is pulled the n-th time, the algorithm observes and collects a reward $X_{k,n}$. $\{X_{k,n}\}_{n\geq 1}$ are i.i.d random variables and we assume that for any k, n, $X_{k,n} - \mu_k$ is 1-sub-Gaussian (The definition can be found for example in Chapter 5 of Lattimore & Szepesvári (2020)). Let X_t denote the reward at time t. For any arm $k \in [K]$, μ_k is its mean reward. $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)$ is the vector of the mean rewards, which is unknown to the algorithm. The largest mean reward is $\mu^* = \max_{k\in[K]} \mu_k$ and the optimal arm is $k^* \in \arg\max_{k\in[K]} \mu_k$. For simplicity of presentation, we assume that the optimal arm is unique. Minor modification can be applied to adapt our results to the non-unique optimal arm case. We define the reward gap such that $\Delta_k = \mu^* - \mu_k \leq \bar{\Delta}, \forall k \neq k^*$, where a known constant $\bar{\Delta} \geq 1$ is an upper bound for all the reward gaps. We also define the relative reward gap $\Delta_{k,k'} = \mu_k - \mu_{k'}, \forall k, k'\ s.t.\ \mu_k \geq \mu_{k'}$. Define the empirical mean for arm k given n observations: $\hat{\mu}_{k,n} = n^{-1} \sum_{k=1}^{n} X_{k,n}$. Let $T_k(t) := \sum_{s=1}^{t} \mathbb{I}\{a_s = k\}$ denote the number of times arm k is pulled from time 1 to time t. As we consider the *Multi-Armed Bandit with Impatient Arms* setup, we introduce a threshold $m_k \in \mathbb{N}^+$ for each arm k. We refer to m_k as arm k's patience. Similar to $\boldsymbol{\mu}$, $\boldsymbol{m} = (m_1, ..., m_K)$ is the vector of $m_k, k \in [K]$. When the algorithm continuously ignores arm k for a duration of m_k time steps, arm k loses patience and exits the game. For example, if $a_{t_0} = k$ and $a_s \neq k$ for all $s = t_0 + 1, ..., t_0 + m_k$, and if $t_0 + m_k + 1 \leq T$, then arm k exits at the end of time $t_0 + m_k$. Arm k can no longer be selected starting from time $t = t_0 + m_k + 1$ as it has already left the game. We assume that all arms are initially full of patience when they enter the game. In other words, if arm k has not been pulled since time $t = 1$, we regard the time when it was last pulled as $t = 0$. We assume that $\boldsymbol{m}$ is given in advance and discuss this in Appendix H. Define a bijective index mapping ind: $[K] \to [K]$ such that $\text{ind}(k)$ is the index of arm k when $\boldsymbol{m}$ is sorted from small to large. Let $\text{ind}^{-1}(i)$ denote the arm whose patience is the i-th small. Following Li et al. (2021), we define the load factor

$$l(K, \boldsymbol{m}) = \sum_{k=1}^{K} \frac{1}{m_k}, \tag{1}$$

as a measure of arm impatience. We adopt cumulative expected regret R_T as the performance metric for algorithms, where

$$R_T = \mathbb{E}\big[\sum_{t=1}^{T} \mu^* - \sum_{t=1}^{T} \mu_{a_t}\big] = \mathbb{E}\big[\sum_{t=1}^{T} \Delta_{a_t}\big]. \tag{2}$$

It is worth noting that, at any time t, the algorithm's choice a_t is compared against k^*, even if k^* may not be available at time t (i.e., $k^* \notin A_t$). This definition of regret is reasonable because the best an algorithm can do is to pull k^* since the beginning and keep it in A_t throughout the time horizon. This reveals one of the crucial differences between our work and the sleeping bandit literature (Gaillard et al., 2023; Kale et al., 2016), whose regrets typically compare a_t against some arm in A_t.

3. Hardness of the Problem and Limitations of the Existing Approaches

3.1. Negative Results in the Low Patience Case

Consider the configuration $(K, \boldsymbol{m}) = (3, (2, 2, 2))$. It can be verified that at least one arm exits at the end of time slot $t = 2$, regardless of which arms the algorithm pulls at times $t = 1$ and 2. Intuitively, this serves as an extreme example where some arms have very low level of patience and no algorithm can guarantee a sub-linear regret under such a $(K, \boldsymbol{m})$ configuration and different $\boldsymbol{\mu}$ values. This is because it is impractical to determine the best arm among the three with only two reward observations. In this section, we formally demonstrate how a low level of patience renders learning infeasible. We first define a quantity characterizing the time when the first exit happens given $(K, \boldsymbol{m})$.

Definition 3.1. Fix $K, \boldsymbol{m}$, define the maximum number of reward observations when the first arm departure happens $Q_{(K,\boldsymbol{m})}$ to be

$$\max\Big\{L \in \mathbb{N} \,\Big|\, \exists \{a_t\}_{t=1}^{L}, \forall s \in [L-1], k \leq K : s < m_k \\ \text{or } s \geq m_k, \exists l = s - m_k + 1, ..., s \text{ } with \text{ } a_l = k\Big\},$$

where $\{a_t\}_{t=1}^{L}$ is a sequence of arm $a_1, a_2, ..., a_L$. If there exists an infinite-length arm sequence such that no arm leaves, let $Q_{(K,\boldsymbol{m})} = +\infty$.

For example, $Q_{(3,(2,2,2))} = 2$. By Definition 3.1, at least one arm exits the game before or at the end of time $t = Q_{(K,\boldsymbol{m})}$, regardless of the algorithm's arm choices. We note that $Q_{(K,\boldsymbol{m})}$ can be uniquely determined by the number of arms K and the vector of patience $\boldsymbol{m}$, although the computation of $Q_{(K,\boldsymbol{m})}$ may be of significant time complexity. Intuitively speaking, the less patient the arms are, the more likely they are to exit early. Thus low level of patience results in a small value of $Q_{(K,\boldsymbol{m})}$. We present a minimax expected regret lower bound depending on $Q_{(K,\boldsymbol{m})}$, which is especially powerful when $Q_{(K,\boldsymbol{m})}$ is small.

Theorem 3.2. *Given a configuration* $(K, \boldsymbol{m})$ *such that* $Q_{(K,\boldsymbol{m})} \leq T$. *Suppose the rewards are independent Gaussian random variables with variance* 1. *Then the minimax expected regret lower bound is given by*

$$\min_{A\in\mathcal{A}} \sup_{\boldsymbol{\mu}\in\Xi} R_T(A, (K, \boldsymbol{m}, \boldsymbol{\mu})) \geq \sqrt{\frac{1}{Q_{(K,\boldsymbol{m})}} \ln \frac{K}{K-1}}$$

$$\times\Big[1 - f_K^{-1}(\frac{1}{2}\ln\frac{K}{K-1})\Big](T - Q_{(K,\boldsymbol{m})}),$$

where $R_T(A, (K, \boldsymbol{m}, \boldsymbol{\mu}))$ is the expected regret when algorithm A is operated on the problem instance $(K, \boldsymbol{m}, \boldsymbol{\mu})$ for a time horizon T. The possible set of algorithms $\mathcal{A} = \big\{\{\pi_t\}_{t=1}^T\big\}$, where $\{\pi_t\}_{t=1}^T$ is a sequence of policies $\pi_1, ..., \pi_T$ that π_t maps from $(a_1, X_1, ..., a_{t-1}, X_{t-1})$ and $\boldsymbol{m}$ to the probability simplex over $[K]$. $\Xi = \{\boldsymbol{\mu} \mid \Delta_k \leq \bar{\Delta}, \forall k \neq k^*\}$. $f_K(p) := h(p) + p\ln\frac{K}{K-1}$ and $h(p) = p\ln p + (1-p)\ln(1-p)$ is the entropy of Bernoulli distribution. $f_K^{-1}(\frac{1}{2}\ln\frac{K}{K-1})$ denotes the unique p such that $f_K(p) = \frac{1}{2}\ln\frac{K}{K-1}$.

Though it can be difficult to compute $Q_{(K,\boldsymbol{m})}$ for arbitrary $K, \boldsymbol{m}$, we find that if the load factor $l(K, \boldsymbol{m}) > 1$, it is possible to obtain upper bounds of $Q_{(K,\boldsymbol{m})}$. In fact, there is a rich set of configurations whose load factor exceeds 1. For instance, $l(3, (2, 2, 2)) = \frac{3}{2} > 1$. We formally describe the upper bound in the following lemma and detail its proof in Appendix C.

Lemma 3.3. *Given a configuration $(K, \boldsymbol{m})$, if $\exists \lambda \in \mathbb{N}^+$ such that $l(K, \boldsymbol{m}) > 1 + \frac{1}{\lambda}$, then we have $Q_{(K,\boldsymbol{m})} \leq \lambda K + 1$.*

Combining Theorem 3.2 and Lemma 3.3, we directly have the following result.

Corollary 3.4. *Fix a configuration $(K, \boldsymbol{m})$ with $l(K, \boldsymbol{m}) > 1$, we have*

$$\min_{A\in\mathcal{A}} \sup_{\boldsymbol{\mu}\in\Xi} R_T(A, (K, \boldsymbol{m}, \boldsymbol{\mu})) = \Omega\big(W(K, \boldsymbol{m})T\big),$$

where $W(K, \boldsymbol{m})$ is a quantity purely depending on $(K, \boldsymbol{m})$. $\mathcal{A}, \Xi$ and $R_T(A, (K, \boldsymbol{m}, \boldsymbol{\mu}))$ are defined in Theorem 3.2.

Corollary 3.4 demonstrates the hardness of algorithm design when $l(K, \boldsymbol{m}) > 1$. If we regard $l(K, \boldsymbol{m})$ as a measure of arm impatience, then $l(K, \boldsymbol{m}) > 1$ is associated with a low level of patience case, when the arms are so likely to leave early that learning their reward distributions is not feasible.

3.2. Analysis of UCB Algorithm

In this part, we study the well-known Upper Confidence Bound (UCB) algorithm in our *Multi-Armed Bandit with Impatient Arms* setup. We adopt the definition in Lattimore & Szepesvári (2020). Define the upper confidence bound of arm k given n observations as $\text{UCB}_k(n) = \hat{\mu}_{k,n} + \sqrt{2n^{-1}\ln\delta^{-1}}$ if $n > 0$ and $\text{UCB}_k(0) = +\infty$. δ is the confidence parameter, typically set to be $\delta = 1/T^2$. The UCB algorithm operates by pulling $a_t = \arg\max_{k\in[K]} \text{UCB}_k(T_k(t))$. We find that when the arms are sufficiently patient, UCB still has performance guarantees. Such results are presented in Appendix D. However, when the optimal arm is impatient, it is highly possible that it exits early under UCB, leaving the algorithm with only sub-optimal arms to choose. We will show that, if there are impatient arms, we can find many problem instances such that UCB performs badly, thus demonstrate its limitation in the proposed setup.

Theorem 3.5. *Run UCB with confidence parameter $\delta = 1/T^2$ in the Multi-Armed Bandit with Impatient Arms setup. Suppose the rewards are independent Gaussian random variables with variance 1. If $\exists\theta \in (0, 1)$, $m_{k^*} = O((\ln T)^\theta)$ and $\exists\beta \neq k^*$ such that $m_\beta \geq T$, then for any $\gamma \in (0, 1)$, we have*

$$R_T = \Omega\Big(\Delta_{\min} T^{1-\gamma}\big(\mathcal{C}(T)\ln T\big)^{-1}\Big),$$

where $\Delta_{\min} := \min_{k\neq k^*}\Delta_k$ and $\mathcal{C}(T)$ is a polynomial function of $\ln T$.

Proof Sketch of Theorem 3.5. We define events $\mathcal{E}_{k^*} = \big\{\min_{n\in[\kappa_T-1]}\text{UCB}_{k^*}(n) > x_1, \text{UCB}_{k^*}(\kappa_T) < x_2\big\}$ and $\mathcal{E}_\beta = \big\{x_2 < \text{UCB}_\beta(a_\beta) < x_1, \min_{n\in[b_\beta]}\text{UCB}_\beta(n) > x_2\big\}$ for arm k^* and β, respectively. $x_1, x_2, \kappa_T, a_\beta, b_\beta$ are constants and $\text{UCB}_k(n)$ denotes the UCB index of arm k given the first n reward observations. We set $x_1 > x_2$ and $a_\beta < b_\beta$. Under these two events, we show that arm k^* is pulled no more than $\kappa_T = o(\ln T)$ times. Since UCB algorithm always pulls the arm with the highest UCB index, arm β is pulled at most a_β times before the κ_T-th pull of arm k^*, given $\min_{n\in[\kappa_T-1]}\text{UCB}_{k^*}(n) > x_1 > \text{UCB}_\beta(a_\beta)$. If arm k^* is pulled the κ_T-th time, at least the $(a_\beta+1)$-th, ..., $(b_\beta+1)$-th pulls of arm β happen before the (κ_T+1)-th pull of arm k^*. If the number of arm β pulls between the κ_T-th and (κ_T+1)-th pull of arm k^* exceeds arm k^*'s patience, then arm k^* leaves the game before it is pulled the (κ_T+1)-th time, not to mention the pulls of other arms. By carefully designing the values of the constants, we show that $\mathcal{E}_{k^*} \cap \mathcal{E}_\beta$ occurs with a non-negligible possibility using some techniques of Brownian motion. □

The complete proof of Theorem 3.5 is also in Appendix D. It shows that as the optimal arm is relatively impatient, the expected regret of UCB algorithm is asymptotically (almost) linear. Fix any $\theta \in (0, 1)$, the patience vector $\boldsymbol{m}$ such that there are negative problem instances for UCB algorithm can have a load factor $l(K, \boldsymbol{m})$ as small as $(\ln T)^{-\theta} + (K-1)/T \to 0$ as $T \to +\infty$, and as large as we wish. If we use load factor as a measure of arm impatience, we see that the capability of UCB algorithm in the proposed setting is very limited, since we can find problem instances such that UCB yields almost linear regret asymptotically even as the load factor approaches 0.

3.3. Analysis of SE Algorithm

In this section, we study the Successive Elimination (SE) algorithm in our *Multi-Armed Bandit with Impatient Arms* setup. We consider a version similar to Algorithm 3 in Even-Dar et al. (2006). All arms are active in the beginning. The algorithm pulls the active arms in a Round-Robin manner, in an increasing order of the arm indices. If at time t all active arms are pulled the same number of times, then at the end of time t the algorithm executes arm elimination when there are more than one active arms. For arm k, if there exists some arm k' such that $\hat{\mu}_{k',T_{k'}(t)} - \hat{\mu}_{k,T_k(t)} > 2\sqrt{4T_k^{-1}(t)\ln T}$, arm k is regarded as sub-optimal and deactivated. Since the arms are scheduled in a Round-Robin fashion, for any arm k, the total number of pulls of other arms between any two consecutive pulls of arm k is at most $K-1$. Then if the patience vector $\boldsymbol{m}$ satisfies that $m_k \geq K, \forall k \in [K]$, no active arm will exit early. Only those deactivated arms may leave the game after being eliminated by the algorithm. The behavior and analysis of SE remain exactly the same in this case as in standard MAB setting, thus, SE is a competitive option when $m_k \geq K, \forall k \in [K]$, i.e., when arms are patient enough in general. However, Round-Robin could induce linear regret when there exists some k such that $m_k < K$. We present this fact in Proposition 3.6 and its proof in Appendix E.

Proposition 3.6. *Run SE in the Multi-Armed Bandit with Impatient Arms setup. Then there exists a problem instance* $(K, \boldsymbol{m}, \boldsymbol{\mu})$ *such that* $l(K, \boldsymbol{m}) < 1$ *but* $R_T = \bar{\Delta}T$.

4. Our Algorithms

From the discussion in Section 3, we know that it is infeasible to design new methods to achieve sub-linear regret when $l(K, \boldsymbol{m})$ is strictly greater than 1. Meanwhile, existing algorithms (e.g., SE) can guarantee sub-linear regret when $m_k \geq K, \forall k \in [K]$. In this section, we aim at developing algorithms for the regime where we do not have clear guarantees, i.e., for $(K, \boldsymbol{m})$ such that $\exists k \in [K], m_k < K$. We notice that in this case, it may be possible to prevent arms' early departure with some carefully designed pulling schedules. We say that a schedule is feasible if no arm leaves early under this schedule. For example, when $\boldsymbol{m} = (2, 4, 4)$, a feasible schedule can be "$a_1, ..., a_6$..." = "1, 2, 1, 3, 1, 2, ...". Formally, define τ_k^i as the time slot when the i-th sample from arm k is scheduled, I_k^i as the time interval (in number of slots) between the i-th and the $(i+1)$-th pull of arm k, $I_k^i = \tau_k^{i+1} - \tau_k^i$. Arm k never leaves the game if $I_k^i \leq m_k, \forall i \geq 1$. Furthermore, we say a schedule "a_1, a_2..." is cyclic if $\exists C \in \mathbb{N}^+$ such that $a_t = a_{t+C}, \forall t \geq 1$. If there exists an infinite-length feasible schedule for $\boldsymbol{m}$, Lemma 4.1 of Li (2023) shows that there must be a feasible cyclic schedule, which consists of repeating finite cycles. We call a finite cycle "$a_1, ..., a_C$" feasible cycle if it forms feasible schedules by repeating itself. If a feasible cycle exists for the considered configuration, it is possible to pull arms according to it to prevent early departure and execute arm elimination at the end of each cycle to remove the sub-optimal arms.

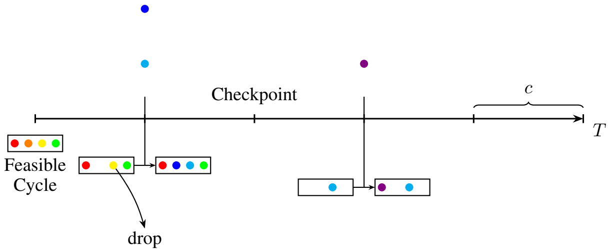

Figure 1. An FC-SE example with $N < K$. In this example, $N = 4, K = 7$. The rectangles denote feasible cycle snapshots. The colored nodes denote different arms.

4.1. FC-SE Algorithm

We introduce the intuition and design details of our Feasible Cycle-based Successive Elimination (FC-SE) algorithm. If a feasible cycle can be found for the considered configuration, we can directly replace the Round-Robin cycle in the original SE algorithm with that feasible cycle. In Section 4.2, we introduce how to construct a feasible cycle. Unfortunately, sometimes a feasible cycle exists only for a subset of the set of all arms. Thus we specify an integer $N \leq K$ such that the subset of arms $\{k : \text{ind}(k) \leq N\}$ can form a feasible cycle. We can first repeat this initial feasible cycle to identify sub-optimal arms and then insert remaining arms into it when some arm is eliminated. However, a problem arises when the cycle is still full as the patience of a remaining arm is about to expire. Since we cannot discard any arm without giving it a chance (it might be the optimal arm), we must remove an arm from the current feasible cycle to make room for an arm that has not been pulled yet. In fact, when a full cycle is repeated many times without any arm being eliminated, it is highly likely that the arms in this cycle have such similar mean rewards that distinguishing a sub-optimal arm becomes difficult. Even if the optimal arm in the current feasible cycle is mistakenly dropped, another arm in the cycle can serve as a suitable substitution. To balance the exploration among arms, we select an appropriate constant c and divide the entire time horizon into segments of equal length c. We tighten the patience of a remaining arm k to $m_k' \leq m_k$, such that m_k' is an integer multiple of c, and pull arm k at around $t = m_k'$ to prevent early departure. We drop arms in the feasible cycle if necessary and add remaining arms to the cycle **only** at these checkpoints (referring to the endpoints of these segments as checkpoints). Figure 1 is a simple example of FC-SE behavior when $N < K$.

We formally present FC-SE in Algorithm 1. Given N, the algorithm constructs a feasible cycle of length $n = \sum_{k:\text{ind}(k)\leq N} n_k$, where n_k is the number of times that arm k appears in the feasible cycle. If $N < K$, we set $n_k = 1$

Algorithm 1 FC-SE

```
Input: Number of arms in the initial feasible cycle $N$, patience vector $\boldsymbol{m}$, time horizon $T$, segment length $c$
Construct a feasible cycle "$\bar{a}_1, ...\bar{a}_n$" for the set of arms $\{k : \text{ind}(k) \le N\}$
$t \leftarrow 1, S \leftarrow \{k : \text{ind}(k) \le N\}, p \leftarrow 1,$ "$\bar{a}'_1, ..., \bar{a}'_n$" $\leftarrow$ "$\bar{a}_1, ..., \bar{a}_n$"
if $N < K$ then
    for $k \in \{k' : \text{ind}(k') > N\}$ do
        $m'_k \leftarrow \left\lceil \frac{\text{ind}(k)-N}{N-1} \right\rceil c$
    end for
end if
while $t \le T$ do
    for $i = 1, ..., n$ do
        if $\bar{a}_i \ne 0$ then
            Pull $a_t = \bar{a}_i$ and receive reward $X_{a_t, T_{a_t}(t)}$
            $t \leftarrow t + 1$
        end if
    end for
    $S \leftarrow \left\{ k \in S : \forall j \in S, \hat{\mu}_{j,T_j(t-1)} - 2\sqrt{\frac{\ln T}{1 \vee T_j(t-1)}} \le \hat{\mu}_{k,T_k(t-1)} + 2\sqrt{\frac{\ln T}{1 \vee T_k(t-1)}} \right\}$
    $\bar{a}_i \leftarrow 0$ for each $i = 1, ..., n$ such that $\bar{a}_i \notin S$
    if $p \le \lceil \frac{K-N}{N-1} \rceil$ and $t + \sum_{k \in S} n_k > cp - n$ then
        $S_{\text{new}} \leftarrow \{k \in [K] : m'_k = cp\}$
        while $|S| + |S_{\text{new}}| > N$ do
            $a \sim \text{Unif}(S)$
            $S \leftarrow S - \{a\}$
        end while
        $\bar{a}_i \leftarrow 0$ for each $i = 1, ..., n$ such that $\bar{a}_i \notin S$
        while $|S_{\text{new}}| > 0$ do
            $a \leftarrow \arg\min_{k \in S_{\text{new}}} \text{ind}(k)$
            $S \leftarrow S \cup \{a\}, S_{\text{new}} \leftarrow S_{\text{new}} - \{a\}$
            $\bar{a}_i \leftarrow a$ where $i = \min\left\{i' \le n : \bar{a}'_{i'} = \min\{k \in [K] \mid \text{ind}(k) \le N \text{ and } \forall j \le n \ s.t. \ \bar{a}'_j = k : \bar{a}_j = 0\}\right\}$
        end while
        $p \leftarrow p + 1$
    end if
end while
```

for any k such that $\text{ind}(k) > N$. The algorithm maintains a set of active arms S, initialized as the set of arms in the initial feasible cycle. p denotes the index of the next checkpoint. The algorithm pulls the active arms according to the feasible cycle. If some arm is eliminated, all positions associated with it in "$\bar{a}_1, ...\bar{a}_n$" are cleared (i.e. set to 0). At the end of each feasible cycle, the algorithm checks whether the next checkpoint is met. It is important to note that there can be at most $N - 1$ remaining arms waiting at a checkpoint, because if more were assigned, we would need to drop all the arms in the current cycle, and there would be no guarantee that a good substitution for the optimal arm remains in the feasible cycle. When checking whether the p-th checkpoint is reached, t is the first time slot of a feasible cycle, which ends at $t + \sum_{k \in S} n_k - 1$. At the first time $t + \sum_{k \in S} n_k > cp - n$, we have that $t \le cp - n$, since t grows at most n each time. When the p-th checkpoint is met, the set of arms that need to be added to the feasible cycle is denoted as S_{new}. Since the actual length of the new feasible cycle is also upper bounded by n, each arm k in S_{new} is pulled no later than $t + n - 1 \le cp - 1 < cp = m'_k \le m_k$. Therefore, arm k does not leave before being pulled by the algorithm for the first time. The algorithm randomly drops arms from the current feasible cycle until there is enough space for S_{new}.

First, we analyse the regret performance of FC-SE algorithm in the easier case when we can find a feasible cycle containing all the available arms (i.e. $N = K$). Note that in this case, the value of c does not matter.

Theorem 4.1. *Run FC-SE in Algorithm 1 with $N = K$. Assume that a feasible cycle can be constructed for the whole arm set $[K]$, with length $n = \sum_{k \ge 1} n_k$. n_k is the number of times that arm k appears in the feasible cycle. Then the expected regret of FC-SE is upper bounded by*

$$R_T \le \sum_{k \ne k^*}^{K} \left[\Delta_k n_k + \frac{32 \ln T}{\Delta_k} \Big(1 + \frac{n_k}{n_{k^*}} \Big) \right] + 2K\Delta_{\max},$$

where $\Delta_{max} := \max_{k \ne k^*} \Delta_k$.

We provide the detailed proof of Theorem 4.1 in Appendix F. Under the assumption of Theorem 4.1, it can be shown that $R_T = O\big(n + \sqrt{nT \ln T}\big)$. Next, we analyse FC-SE in the case when we specify some $N < K$ such that a feasible cycle exists for the subset of arms $\{k : \text{ind}(k) \le N\}$. The proof is also in Appendix F.

Theorem 4.2. *Run FC-SE in Algorithm 1 with $N < K$. Assume that a feasible cycle can be constructed for the set of arms $\{k : ind(k) \le N\}$, with length $n = \sum_{k: ind(k) \le N} n_k$ and $n_k = 1$ for any k such that $ind(k) > N$. We set c to be*

$$\min \left\{ \left\lfloor \min_{k: ind(k) > N} \frac{m_k}{\lceil \frac{ind(k) - N}{N-1} \rceil} \right\rfloor, \right. \tag{3}$$

$$\left. 3n + \left\lceil \left(\frac{4T\sqrt{n \ln T}}{(K-1)\bar{\Delta}} \right)^{\frac{2}{3}} \right\rceil \right\}. \tag{4}$$

If we have $\left\lfloor \min_{k: ind(k) > N} m_k / \lceil \frac{ind(k)-N}{N-1} \rceil \right\rfloor > 3n$, then the expected regret of FC-SE is upper bounded as

$$R_T \le (K-1)\bar{\Delta} c \Big(1 + \Big\lceil \frac{K-N}{N-1} \Big\rceil \Big) + 8T \Big(1 + \Big\lceil \frac{K-N}{N-1} \Big\rceil \Big) \sqrt{\frac{n \ln T}{c - 3n}} + 2K\Delta_{\max},$$

Algorithm 2 The Shortest Length AUS Cycle Construction

```
Input: Patience vector $\boldsymbol{m}$
Compute the index mapping $\text{ind}(\cdot)$ such that $m_{\text{ind}^{-1}(1)} \leq m_{\text{ind}^{-1}(2)} \leq ... \leq m_{\text{ind}^{-1}(K)}$
Solve the optimization problem and obtain solution $\boldsymbol{r}^*$:
    $\min_{\hat{\boldsymbol{r}}} \quad \frac{\sum_{k=1}^{K} \hat{r}_k}{\hat{r}_K}$   (5)
    $s.t. \quad \frac{\hat{r}_k}{\hat{r}_{k+1}} \in \mathbb{N}^+, \ \forall k \in [K-1]$   (6)
    $\frac{1}{m_{\text{ind}^{-1}(k)}} \leq \hat{r}_k \leq 1, \ \forall k \in [K]$   (7)
    $\sum_{k=1}^{K} \hat{r}_k \leq 1$   (8)
if $\boldsymbol{r}^*$ exists then
    Construct an AUS cycle with $\boldsymbol{r}^*$
end if
```

where $\Delta_{max} := \max_{k \neq k^*} \Delta_k$. *Specifically, if* $\boldsymbol{m}$ *satisfies that* $\left\lfloor \min_{k:ind(k)>N} m_k / \left\lceil \frac{ind(k)-N}{N-1} \right\rceil \right\rfloor > 3n + \left\lceil \left(\frac{4T\sqrt{n \ln T}}{(K-1)\bar{\Delta}} \right)^{\frac{2}{3}} \right\rceil$, *we have* $R_T = O\big(K^{\frac{4}{3}} T^{\frac{2}{3}} (n \ln T)^{\frac{1}{3}}\big)$.

4.2. A Form of Feasible Cycles: AUS

In Section 4.1, we presented our first algorithm, FC-SE. This algorithm requires a feasible cycle containing at least a subset of all the available arms. Besides, the expected regret increases with the feasible cycle length n in our analysis. In this section, we propose a specific method to construct a feasible cycle with the shortest possible length. We need the following definition from the Age of Information (AoI) literature.

Definition 4.3 (Li et al. (2023a)). A cyclic schedule is an Almost Uniform Schedule (AUS) if for each arm k, there exists an integer x_k such that I_k^i is either x_k or x_k+1 for any $i \geq 1$. I_k^i is the time interval (in number of slots) between the i-th and the $(i+1)$-th pull of arm k in the cyclic schedule.

Our process to construct a feasible cycle is listed in Algorithm 2. Given any feasible solution to the optimization problem in Algorithm 2, a routine to construct a feasible AUS cycle is described in Li et al. (2021). Besides, our minimization objective (5) is the length of the constructed AUS cycle. In conclusion, Algorithm 2 tries to find the feasible AUS cycle with the shortest possible length. In Appendix B.1, we present a dynamic programming-based solution to the optimization problem in Algorithm 2. Here we formally describe the output of Algorithm 2 and provide the proof in Appendix B.2.

Theorem 4.4. *Run Algorithm 2 on a patience vector* $\boldsymbol{m}$. *If the optimization problem in Algorithm 2 has a feasible solution, then Algorithm 2 finds a feasible AUS cycle of length at most* $||\boldsymbol{m}||_\infty$.

Remark 4.5. Although it is obvious that when $l(K, \boldsymbol{m}) > 1$, no solution to the optimization problem exists, Li et al. (2021) has shown that $l(K, \boldsymbol{m}) \leq \ln 2$ can ensure the existence of such a feasible solution. For some configurations that $\min_{k \in [K]} m_k < K$ but $l(K, \boldsymbol{m}) \leq \ln 2$, though it cannot be handled by the SE algorithm, it is possible to schedule the arms such that no arm leaves early. For instance, in many configurations where $||\boldsymbol{m}||_\infty = O(\ln T)$, it is also highly likely that $l(K, \boldsymbol{m})$ falls below the constant $\ln 2$ as T becomes sufficiently large, making feasible AUS cycles exist. As Theorem 4.4 indicates, Algorithm 2 finds AUS cycles of length $n = O(\ln T)$ for such configurations. As a consequence, we can find a wide range of configurations results in AUS cycles of sub-linear length, leading to the sub-linear regret of our novel algorithms.

4.3. Extension: Allowing New Entering Arms

In the proposed setup, we only assume that arms can exit. In reality, a more general and realistic scenario involves both incoming and departing arms. In online advertising, as time progresses, more advertisers may attempt to display their advertisements. Similarly, in the crowdsourcing example, new workers might join the platform and search for jobs. In this section, we take a step forward and allow for the entrance of new arms in the proposed setting. There are also a total of K arms, but only the set of arms $[K_0]$ is available at the beginning, where $K_0 < K$. The remaining arms $K_0 + 1, ..., K$ arrive later within the time horizon T. ρ_k denotes the time slot at the beginning of which arm $k \in [K]$ enters the game. We have $\rho_k = 0, \forall k \in [K_0]$. Without loss of generality, we assume that arms $k = K_0 + 1, ..., K$ are ordered by their entry times: $0 < \rho_{K_0+1} \leq ... \leq \rho_K$. For any $k > K_0$, we assume that the entry time ρ_k and the patience m_k is not known in advance. The algorithm does not know m_k and when arm k becomes available until the beginning of time ρ_k. In this section, we define the mapping $\text{ind}(\cdot)$ only for the set of initial arms $[K_0]$: $m_{\text{ind}^{-1}(1)} \leq ... \leq m_{\text{ind}^{-1}(K_0)}$. When designing performance metrics for algorithms in this new setup, we notice that it is possible for the optimal arm k^* to be among the new entering arms (i.e. $k^* > K_0$). In this case, from $t = 1$ to $t = \rho_{k^*} - 1$, no matter which arm we pull, there is a positive gap in reward mean when comparing a_t against arm k^*. Even the Oracle that always pulls the best possible arm yields a positive regret that is linear with respect to ρ_{k^*}. The expected regret (2) is no longer suitable. Instead, we introduce the expected dynamic regret

$$\tilde{R}_T = \mathbb{E}\big[\sum_{t=1}^{T} \mu_t^* - \sum_{t=1}^{T} \mu_{a_t}\big], \tag{9}$$

where $\mu_t^* := \max_{k\in[K]:\rho_k\le t} \mu_k$ is the highest reward mean of the arms that have entered the game before or at time t.

To handle newly entering arms with unknown patience, we reserve special slots in the feasible cycle. In the beginning, we use only $N-2$ of the initially available arms with relatively low level of patience and introduce two additional virtual arms to construct the feasible cycle, where N is also the maximum number of arms in the cycle. Although m_k is unknown before arm k's entrance, we assume a known lower bound $\underline{m}$ for the patience of newly entering arms. The two virtual arms are denoted as $+$ and $-$, and we set their patience as $m_+ = m_- = \underline{m}$. The reserved slots for the virtual arms in the feasible cycle are initially empty. New entering arms occupy the slots of either $+$ or $-$ after their entrance. The condition $\underline{m} \le m_k$ ensures that the slots of either virtual arm are sufficient to keep any entering arm. The details of FC-Entry is in Appendix G.

Theorem 4.6. *Run FC-Entry algorithm in Algorithm 4 with N. Assume N satisfies that $3 < N < K_0 + 2$, the set of arms $\{k \in [K_0] : ind(k) \le N-2\}$ and two virtual arms $+, -$ with patience $\underline{m}$ can form a feasible cycle of length $n = n_+ + n_- + \sum_{k\in[K_0]:ind(k)\le N-2} n_k$. n_+, n_- are the numbers of pulls of the virtual arms $+, -$ in the constructed feasible cycle, respectively. $n_k = 1$ for $k \in [K_0] : ind(k) > N-2$. For arm $k > K_0$, $n_k = n_+$ if it takes up the slots of virtual arm $+$ in the feasible cycle and otherwise $n_k = n_-$. We set c to be*

$$\min\Bigg\{\Bigg\lfloor \min_{k\in[K_0]:ind(k)>N-2} \frac{m_k}{\lceil\frac{ind(k)-N+2}{N-3}\rceil}\Bigg\rfloor, \tag{10}$$

$$3n + \Bigg\lceil\Bigg(\frac{8(K-N+3)T\sqrt{n\ln T}}{\bar{\Delta}(3+\lceil\frac{K_0-N+2}{N-3}\rceil)}\Bigg)^{\frac{2}{3}}\Bigg\rceil\Bigg\}. \tag{11}$$

If $c > 3n$ and $3c \le \min_{K_0<k<K}(\rho_{k+1} - \rho_k - 1)$ then the expected dynamic regret of FC-Entry is upper bounded as

$$\tilde{R}_T \le 2K\Delta_{\max} + (K-1)\bar{\Delta}c\Bigg(3 + \Bigg\lceil\frac{K_0-N+2}{N-3}\Bigg\rceil\Bigg) + 16(K-1)(K-N+3)T\sqrt{\frac{n\ln T}{c-3n}},$$

where $\Delta_{max} := \max_{k\ne k^*}\Delta_k$. If c is exactly $3n + \lceil\big(\frac{8(K-N+3)T\sqrt{n\ln T}}{\bar{\Delta}(3+\lceil\frac{K_0-N+2}{N-3}\rceil)}\big)^{\frac{2}{3}}\rceil$, $\tilde{R}_T = O\big(K^2T^{\frac{2}{3}}(n\ln T)^{\frac{1}{3}}\big)$.

5. Numerical Experiments

We examine the theoretical results in 4 simulations. In a simulation, each curve is the average over 50 trials with i.i.d. standard Gaussian noises. $T = 10^5$ in each simulation.

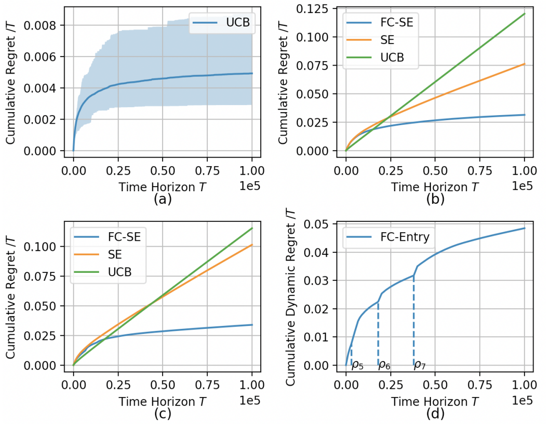

Figure 2. Simulation results. (a) UCB with sufficient arm patience; (b) FC-SE when $N = K$; (c) FC-SE when $N < K$; (d) FC-Entry with newly entering arms. The three new arms $5, 6, 7$ enter the game at time $\rho_5 = 3000, \rho_6 = 18000, \rho_7 = 38000$.

UCB. We consider there are $K = 5$ arms. Let $\mu_1 = 0.7$. Other reward means are sampled uniformly from $[0, 0.6]$, thus $k^* = 1$ and $\boldsymbol{\mu} \in \Xi_\epsilon$ (defined in Corollary D.2) with $\epsilon = 0.1$ and $\bar{\Delta} = 0.7$. The entries of $\boldsymbol{m}$ are sampled uniformly from $\big[1 + (K-1)\lceil 16\epsilon^{-2}\ln T\rceil, T\big]$. We run UCB with $\delta = 1/T^2$ to validate the sub-linear regret upper bound in Appendix D. Fig. 2(a) shows the obtained regret curve. The figure also shows the realized minimum regret and maximum regret in the shaded area.

FC-SE with $N = K$. We consider there are $K = 5$ arms and set $\boldsymbol{m} = (3, 5, 12, 155, 1000)$. We run Algorithm 2 and construct a feasible cycle ”1, 2, 3, 1, 2, 4, 1, 2, 5” with $n = 9$ for all the arms. The entries of $\boldsymbol{\mu}$ are sampled uniformly from $[0, 1]$. We run UCB (with $\delta = 1/T^2$) and SE as baseline algorithms for FC-SE. In each run of FC-SE, no arm exits the game before it is regarded as sub-optimal. Fig. 2(b) shows the obtained regret curves.

FC-SE with $N < K$. We consider there are $K = 6$ arms and set $\boldsymbol{m} = (2, 4, 4, 6800, 6800, 15000)$. We let $N = 3$ and use the feasible cycle ”1, 2, 1, 3” with $n = 4$. The entries of $\boldsymbol{\mu}$ are sampled uniformly from $[0, 1]$. We run UCB (with $\delta = 1/T^2$) and SE as baseline algorithms. In each run of FC-SE, no arm exits the game before being regarded as sub-optimal. Fig. 2(c) shows the obtained regret curves.

FC-Entry. We consider there are $K = 7$ arms and set $\boldsymbol{m} = (3, 5, 1000, 6667, 12, 10000, 26)$. Arm $5, 6, 7$ are newly entering arms and $\underline{m} = 12$. We construct a feasible cycle ”1, 2, +, 1, 2, -, 1, 2, 3” with $n = 9$ for patience vector $(m_1, m_2, m_+, m_-, m_3) = (3, 5, 12, 12, 1000)$, where $+, -$ are two virtual arms with patience $\underline{m}$. The entries of $\boldsymbol{\mu}$ are sampled uniformly from $[0, 1]$. In each run of FC-Entry, no arm exits the game before it is regarded as sub-optimal. Fig. 2(d) shows the obtained dynamic regret curve.

Acknowledgements

The authors thank the anonymous reviewers for their helpful comments. This work is supported by Tsinghua University Dushi Program.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbas, Q., Zeb, S., Hassan, S. A., Mumtaz, R., and Zaidi, S. A. R. Joint optimization of age of information and energy efficiency in iot networks. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pp. 1–5, 2020. doi: 10.1109/VTC2020-Spring48590.2020.9129207.

Abbas, Q., Hassan, S. A., Qureshi, H. K., Dev, K., and Jung, H. A comprehensive survey on age of information in massive iot networks. *Computer Communications*, 197:199–213, 2023. ISSN 0140-3664. doi: https://doi.org/10.1016/j.comcom.2022.10.018. URL https://www.sciencedirect.com/science/article/pii/S0140366422004066.

Abhishek, K., Ghalme, G., Gujar, S., and Narahari, Y. Sleeping combinatorial bandits. *CoRR*, abs/2106.01624, 2021. URL https://arxiv.org/abs/2106.01624.

Agarwal, A., Khanna, S., and Patil, P. A sharp memory-regret trade-off for multi-pass streaming bandits. In Loh, P.-L. and Raginsky, M. (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 1423–1462. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/agarwal22a.html.

Arafa, A., Yang, J., Ulukus, S., and Poor, H. V. Age-minimal transmission for energy harvesting sensors with finite batteries: Online policies. *IEEE Transactions on Information Theory*, 66(1):534–556, 2020. doi: 10.1109/TIT.2019.2938969.

Aramayo, N., Schiappacasse, M., and Goic, M. A multi-armed bandit approach for house ads recommendations. *Marketing Science*, 42(2):271–292, 2023.

Assadi, S. and Wang, C. Single-pass streaming lower bounds for multi-armed bandits exploration with instance-sensitive sample complexity. *Advances in Neural Information Processing Systems*, 35:33066–33079, 2022.

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

Avadhanula, V., Colini Baldeschi, R., Leonardi, S., Sankararaman, K. A., and Schrijvers, O. Stochastic bandits for multi-platform budget optimization in online advertising. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 2805–2817, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450074. URL https://doi.org/10.1145/3442381.3450074.

Aziz, M., Kaufmann, E., and Riviere, M.-K. On multi-armed bandit designs for dose-finding clinical trials. *J. Mach. Learn. Res.*, 22(1), jan 2021. ISSN 1532-4435.

Basu, S., Sankararaman, K. A., and Sankararaman, A. Beyond $log^2(t)$ regret for decentralized bandits in matching markets. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 705–715. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/basu21a.html.

Chakrabarti, D., Kumar, R., Radlinski, F., and Upfal, E. Mortal multi-armed bandits. *Advances in neural information processing systems*, 21, 2008.

Chakravorty, J. and Mahajan, A. Multi-armed bandits, gittins index, and its calculation. *Methods and applications of statistics in clinical trials: Planning, analysis, and inferential methods*, 2:416–435, 2014.

Chaudhuri, A. R. and Kalyanakrishnan, S. Regret minimisation in multi-armed bandits using bounded arm memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10085–10092, 2020.

Chen, S., Yang, N., Zhang, M., and Wang, J. Minimizing age of information for mobile edge computing systems: A nested index approach, 2023.

Cortes, C., Desalvo, G., Gentile, C., Mohri, M., and Yang, S. Online learning with sleeping experts and feedback graphs. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1370–1378. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/cortes19a.html.

Duembgen, L. Bounding standard gaussian tail probabilities. *arXiv preprint arXiv:1012.2063*, 2010.

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006. URL http://jmlr.org/papers/v7/evendar06a.html.

Feki, A. and Capdevielle, V. Autonomous resource allocation for dense lte networks: A multi armed bandit formulation. In *2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 66–70, 2011. doi: 10.1109/PIMRC.2011.6140047.

Gaillard, P., Saha, A., and Dan, S. One arrow, two kills: A unified framework for achieving optimal regret guarantees in sleeping bandits. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7755–7773. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/gaillard23a.html.

Han, B. and Gabor, J. Contextual bandits for advertising budget allocation. *Proceedings of the ADKDD*, 17, 2020.

Huang, Z., Xu, Y., Hu, B., Wang, Q., and Pan, J. Thompson sampling for combinatorial semi-bandits with sleeping arms and long-term fairness constraints. *CoRR*, abs/2005.06725, 2020. URL https://arxiv.org/abs/2005.06725.

Jin, T., Huang, K., Tang, J., and Xiao, X. Optimal streaming algorithms for multi-armed bandits. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5045–5054. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jin21a.html.

Kale, S., Lee, C., and Pal, D. Hardness of online sleeping combinatorial optimization problems. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/184260348236f9554fe9375772ff966e-Paper.pdf.

Kanade, V. and Steinke, T. Learning hurdles for sleeping experts. *ACM Trans. Comput. Theory*, 6(3), jul 2014. ISSN 1942-3454. doi: 10.1145/2505983. URL https://doi.org/10.1145/2505983.

Kanade, V., McMahan, H. B., and Bryan, B. Sleeping experts and bandits with stochastic action availability and adversarial rewards. In van Dyk, D. and Welling, M. (eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 272–279, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL https://proceedings.mlr.press/v5/kanade09a.html.

Kaul, S., Gruteser, M., Rai, V., and Kenney, J. Minimizing age of information in vehicular networks. In *2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, pp. 350–358, 2011. doi: 10.1109/SAHCN.2011.5984917.

Kaul, S., Yates, R., and Gruteser, M. Real-time status: How often should one update? In *2012 Proceedings IEEE INFOCOM*, pp. 2731–2735, 2012. doi: 10.1109/INFCOM.2012.6195689.

Kong, F., Yin, J., and Li, S. Thompson sampling for bandit learning in matching markets. In Raedt, L. D. (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 3164–3170. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/439. URL https://doi.org/10.24963/ijcai.2022/439. Main Track.

Lai, T. L. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, 15(3):1091–1114, 1987. ISSN 00905364. URL http://www.jstor.org/stable/2241818.

Lancewicki, T., Segal, S., Koren, T., and Mansour, Y. Stochastic multi-armed bandits with unrestricted delay distributions. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5969–5978. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/lancewicki21a.html.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Li, C. *Optimizing Information Freshness in Wireless Networks*. PhD thesis, Virginia Tech, 2023.

Li, C., Liu, Q., Li, S., Chen, Y., Hou, Y. T., and Lou, W. On scheduling with aoi violation tolerance. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pp. 1–9, 2021. doi: 10.1109/INFOCOM42981.2021.9488685.

Li, C., Li, S., Liu, Q., Hou, Y. T., Lou, W., and Kompella, S. Eywa: A general approach for scheduler design in aoi optimization. In *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, pp. 1–9, 2023a. doi: 10.1109/INFOCOM53939.2023.10228973.

Li, F., Liu, J., and Ji, B. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3):1799–1813, 2020. doi: 10.1109/TNSE.2019.2954310.

Li, S., Zhang, L., Wang, J., and Li, X.-Y. Tight memory-regret lower bounds for streaming bandits, 2023b.

Liau, D., Song, Z., Price, E., and Yang, G. Stochastic multi-armed bandits in constant space. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 386–394. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/liau18a.html.

Ling, Z., Hu, F., Zhang, H., and Han, Z. Age-of-information minimization in healthcare iot using distributionally robust optimization. *IEEE Internet of Things Journal*, 9(17): 16154–16167, 2022. doi: 10.1109/JIOT.2022.3150321.

Liu, L. T., Mania, H., and Jordan, M. Competing bandits in matching markets. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1618–1628. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/liu20c.html.

Liu, L. T., Ruan, F., Mania, H., and Jordan, M. I. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22(211):1–34, 2021a. URL http://jmlr.org/papers/v22/20-1429.html.

Liu, Q., Zeng, H., and Chen, M. Minimizing age-of-information with throughput requirements in multi-path network communication. In *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Mobihoc '19, pp. 41–50, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367646. doi: 10.1145/3323679.3326502. URL https://doi.org/10.1145/3323679.3326502.

Liu, Q., Li, C., Hou, Y. T., Lou, W., and Kompella, S. Aion: A bandwidth optimized scheduler with aoi guarantee. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pp. 1–10, 2021b. doi: 10.1109/INFOCOM42981.2021.9488781.

Liu, Q., Li, C., Hou, Y. T., Lou, W., Reed, J. H., and Kompella, S. Ao2i: Minimizing age of outdated information to improve freshness in data collection. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, pp. 1359–1368, 2022. doi: 10.1109/INFOCOM48880.2022.9796932.

Liu, Y. and Liu, M. An online learning approach to improving the quality of crowd-sourcing. *IEEE/ACM Transactions on Networking*, 25(4):2166–2179, 2017. doi: 10.1109/TNET.2017.2680245.

Lou, A. and Goldfeld, Z. Ece 5630 lecture 7: Data processing inequality, February 2020.

Lu, N., Ji, B., and Li, B. Age-based scheduling: Improving data freshness for wireless real-time traffic. In *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Mobihoc '18, pp. 191–200, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357708. doi: 10.1145/3209582.3209602. URL https://doi.org/10.1145/3209582.3209602.

Mahadik, K., Wu, Q., Li, S., and Sabne, A. Fast distributed bandits for online recommendation systems. In *Proceedings of the 34th ACM International Conference on Supercomputing*, ICS '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379830. doi: 10.1145/3392717.3392748. URL https://doi.org/10.1145/3392717.3392748.

Maiti, A., Patil, V., and Khan, A. Multi-armed bandits with bounded arm-memory: Near-optimal guarantees for best-arm identification and regret minimization. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19553–19565. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a2f04745390fd6897d09772b2cd1f581-Paper.pdf.

Mehta, A. et al. Online matching and ad allocation. *Foundations and Trends® in Theoretical Computer Science*, 8 (4):265–368, 2013.

Ni, Y., Cai, L., and Bo, Y. Vehicular beacon broadcast scheduling based on age of information (aoi). *China Communications*, 15(7):67–76, 2018. doi: 10.1109/CC.2018.8424604.

Nika, A., Elahi, S., and Tekin, C. Contextual combinatorial volatile multi-armed bandit with adaptive discretization. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial*

Intelligence and Statistics, volume 108 of *Proceedings of Machine Learning Research*, pp. 1486–1496. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/nika20a.html.

Pase, F., Giordani, M., Cuozzo, G., Cavallero, S., Eichinger, J., Verdone, R., and Zorzi, M. Distributed resource allocation for urllc in iiot scenarios: A multi-armed bandit approach. In *2022 IEEE Globecom Workshops (GC Wkshps)*, pp. 383–388, 2022. doi: 10.1109/GCWkshps56602.2022.10008671.

Qin, Z., Yang, S., Huang, Y., Fu, H., Zhou, P., and Ding, G. Towards relevance and diversity in crowdsourcing worker recruitment with insufficient information. *IEEE Transactions on Network Science and Engineering*, pp. 1–14, 2023. doi: 10.1109/TNSE.2023.3302375.

Rangi, A. and Franceschetti, M. Multi-armed bandit algorithms for crowdsourcing systems with online estimation of workers' ability. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pp. 1345–1352, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

Rathod, S. On reducing the order of arm-passes bandit streaming algorithms under memory bottleneck, 2021.

Rigollet, P. and Hütter, J.-C. High dimensional statistics. *Lecture notes for course 18.657 at MIT*, 2019.

Robbins, H. E. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1952. URL https://api.semanticscholar.org/CorpusID:15556973.

Saha, A., Gaillard, P., and Valko, M. Improved sleeping bandits with stochastic actions sets and adversarial rewards. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Sankararaman, A., Basu, S., and Abinav Sankararaman, K. Dominate or delete: Decentralized competing bandits in serial dictatorship. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1252–1260. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/sankararaman21a.html.

Santana, M. R. O., Melo, L. C., Camargo, F. H. F., Brandão, B., Soares, A., Oliveira, R. M., and Caetano, S. Contextual meta-bandit for recommender systems selection. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, pp. 444–449, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832. doi: 10.1145/3383313.3412209. URL https://doi.org/10.1145/3383313.3412209.

Scheike, T. H. A boundary-crossing result for brownian motion. *Journal of Applied Probability*, 29(2):448–453, 1992.

Schwartz, E. M., Bradlow, E. T., and Fader, P. S. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4): 500–522, 2017.

Song, Y. and Jin, H. Minimizing entropy for crowdsourcing with combinatorial multi-armed bandit. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pp. 1–10, 2021. doi: 10.1109/INFOCOM42981.2021.9488800.

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

Tracà, S., Rudin, C., and Yan, W. Reducing exploration of dying arms in mortal bandits. In Adams, R. P. and Gogate, V. (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 156–163. PMLR, 22–25 Jul 2020. URL https://proceedings.mlr.press/v115/traca20a.html.

Tran-Thanh, L., Stein, S., Rogers, A., and Jennings, N. R. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence*, 214:89–111, 2014. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2014.04.005. URL https://www.sciencedirect.com/science/article/pii/S0004370214000538.

Wang, C. Tight regret bounds for single-pass streaming multi-armed bandits. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35525–35547. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/wang23a.html.

Xie, Z., Yu, T., Zhao, C., and Li, S. Comparison-based conversational recommender system with relative bandit feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 1400–1409, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462920. URL https://doi.org/10.1145/3404835.3462920.

Yang, H. and Lu, Q. Dynamic contextual multi arm bandits in display advertisement. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 1305–1310, 2016. doi: 10.1109/ICDM.2016.0177.

Yates, R. D., Sun, Y., Brown, D. R., Kaul, S. K., Modiano, E., and Ulukus, S. Age of information: An introduction and survey. *IEEE Journal on Selected Areas in Communications*, 39(5):1183–1210, 2021. doi: 10.1109/JSAC.2021.3065072.

Zhang, H., Ma, Y., and Sugiyama, M. Bandit-based task assignment for heterogeneous crowdsourcing. *Neural Computation*, 27(11):2447–2475, 2015. doi: 10.1162/NECO_a_00782.

Zhang, Y., Wang, S., and Fang, Z. Matching in multi-arm bandit with collision. *Advances in Neural Information Processing Systems*, 35:9552–9563, 2022.

Zuo, J. and Joe-Wong, C. Combinatorial multi-armed bandits for resource allocation. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–4, 2021. doi: 10.1109/CISS50987.2021.9400228.

A. Details of Related Work

A.1. Bandit with Time-Varying Arm Availability

In many real world applications, the available arm set usually varies over time (Saha et al., 2020). In addition to the sleeping bandit literature we introduced earlier, there is research on combinatorial bandits with sleeping arms (Kale et al., 2016; Nika et al., 2020; Huang et al., 2020; Abhishek et al., 2021; Li et al., 2020), where they permit the set of base arms to evolve. As we have mentioned in the main body, one of the key differences of our work from this literature is the difference of suitable performance metrics. In the literature, the selected action is compared with some arm in A_t. To illustrate the potential pitfalls of such performance metrics, consider an example where the patience of each arm is relatively low and a simple algorithm that keeps pulling some fixed sub-optimal arm k. As the other arms have all departed, k becomes the only available option. The regret ceases to increase as arm k is only compared against itself. While this simple algorithm is satisfactory in terms of the regret defined in these papers, it is important to note that it incurs a linear loss with respect to the time horizon T if compared to an algorithm that consistently pulls the optimal arm. The appropriate performance metric is then to compare the algorithm choice against the optimal arm that has ever appeared, regardless of whether it is still available at time t. In fact, a well-designed algorithm should aim to prevent the unexpected departure of the optimal arm, as its early exit can result in significant regret. Besides, we note that a concept of patience is also introduced in (Mehta et al., 2013). The patience there represents the allowed maximum times of pulls for each arm. Say each arm is associated with a counter recording the number of times it has been pulled. The value in the counter is non-decreasing and the counter does not reset when the arm is ignored in some round. In contrast, the patience defined in this paper is a threshold for the time an arm is continuously ignored. Say each arm has a counter recording the time that has elapsed since the last pull of it. The counter is reset once the arm is pulled. This is why it is possible to schedule some arms infinitely without ever violating our patience thresholds.

A.2. Streaming Bandit

Recently, motivated by the fact that the number of bandit arms can be notably extensive while the learner often contends with limited available memory, a line of research (Liau et al., 2018; Chaudhuri & Kalyanakrishnan, 2020; Maiti et al., 2021; Rathod, 2021; Jin et al., 2021; Assadi & Wang, 2022; Agarwal et al., 2022; Wang, 2023; Li et al., 2023b) have defined and studied the Streaming Bandit setting. In this context, the arms arrive one at a time in a sequential manner, and the algorithm is constrained to pulling arms currently stored in memory. The size of the memory is generally much smaller than the total number of arms. When the memory becomes saturated and the algorithm aims to select a new arm, it is necessary to remove at least one arm already stored in memory before the new selection can be made (Li et al., 2023b). This model is relevant to ours for two reasons. Firstly, while our setting does not consider bounded memory, an algorithm must choose to operate on only a subset of arms when the arms have limited patience. This is necessary because attempting to manage all arms may deplete the patience of some. Secondly, both this model and ours allow for the entry and exit of arms within the game process. Although, in most parts of our work, we assume that there are no arm entries. The setup of streaming bandits can be categorized into two distinct types. The first is the single-pass setting (Maiti et al., 2021; Assadi & Wang, 2022; Wang, 2023), in which the algorithm is unable to re-add an arm to the arm memory that was previously removed from it. The second is the multi-pass setting (Liau et al., 2018; Chaudhuri & Kalyanakrishnan, 2020; Rathod, 2021; Agarwal et al., 2022), where the algorithm is permitted to scan the stream of arms a limited number of times, thus allowing for the retrieval of arms that were previously discarded. Compared to the multi-pass streaming bandit scenario, the setup we investigate in this paper is more akin to the single-pass streaming bandit configuration, primarily because arms that have exited cannot be recovered by the algorithm. Notably, our model differs from the streaming bandit framework by assuming that arms possess the capacity to autonomously determine whether to persist in participation or withdraw from the game. Despite the algorithm potentially being aware of their exit strategy, arms retain a certain degree of decision-making power. In contrast, in the streaming bandit, it is the algorithm that dictates whether to retain an arm in memory or to discard it.

A.3. Age of Information (AoI)

The Age of Information (AoI) is a metric within the application layer of computer networks that characterizes the freshness of information. In its basic model, there are multiple data collection sources and a base station. Each source collects its data samples and transmits them to the base station through a shared wireless channel (Liu et al., 2022). AoI is defined as the time elapsed since the generation of the latest received message at the base station. Since its original introduction in Kaul et al. (2011; 2012), AoI has recently garnered significant research attention. Several survey papers (Yates et al., 2021; Abbas et al., 2023) offer comprehensive summaries of the latest advancements in AoI research. AoI is particularly suitable for scenarios where the timeliness of status updates is critical, and as a result, it finds a wide range of applications in intelligent transportation systems (Ni et al., 2018), cellular-based IoT systems (Ling et al., 2022), wireless ad hoc network traffic scheduling (Lu et al., 2018), smart agriculture (Abbas et al., 2020), and so on.

B. Details of AUS construction

B.1. Solution to the Optimization Problem in Algorithm 2

We design a dynamic programming-based algorithm to compute the solution to the optimization problem in Algorithm 2. The details are presented in Algorithm 3. The algorithm is based on a crucial observation: say $\hat{\boldsymbol{r}}$ is a feasible solution to the optimization problem in Algorithm 2, then it can be mapped to another feasible solution $\hat{\boldsymbol{r}}'$ such that $\exists k \in [K] : \hat{r}'_k = \frac{1}{m_{\text{ind}^{-1}(k)}}$ and $\frac{\sum_{k=1}^K \hat{r}_k}{\hat{r}_K} = \frac{\sum_{k=1}^K \hat{r}'_k}{\hat{r}'_K}$. If $\hat{\boldsymbol{r}}$ itself satisfies that $\exists k \in [K] : \hat{r}_k = \frac{1}{m_{\text{ind}^{-1}(k)}}$, we can simply set $\hat{\boldsymbol{r}}' = \hat{\boldsymbol{r}}$. Otherwise we can set $\hat{\boldsymbol{r}}' = a\hat{\boldsymbol{r}}$ where $a = \max_{k' \in [K]} \frac{1}{\hat{r}_{k'} m_{\text{ind}^{-1}(k')}} < 1$. It can be verified that $\hat{\boldsymbol{r}}'$ is also a feasible solution and $\frac{\sum_{k=1}^K \hat{r}'_k}{\hat{r}'_K} = \frac{\sum_{k=1}^K a\hat{r}_k}{a\hat{r}_K} = \frac{\sum_{k=1}^K \hat{r}_k}{\hat{r}_K}$. Given this fact, it suffices to find an optimal $\hat{\boldsymbol{r}}$ such that $\exists k \in [K] : \hat{r}_k = \frac{1}{m_{\text{ind}^{-1}(k)}}$. In the outer iteration, the algorithm iterates over k and fixes $\hat{r}_k = \frac{1}{m_{\text{ind}^{-1}(k)}}$. Given k, for $s \leq k$, DP(r, s) denotes the minimum value of $\sum_{k'=s}^k \hat{r}_{k'}$ where $\hat{r}_s = r$ and $\hat{r}_s, ..., \hat{r}_k$ satisfy (6), (7). Similarly, for $s > k$, the sum is from k to s. For $s \leq k$, BWD(r, s) denotes the $\hat{r}_{s+1}$ that achieves DP(r, s), while for $s > k$, BWD(r, s) denotes the $\hat{r}_{s-1}$ that achieves DP(r, s). If $k > 1$, the algorithm minimizes $\sum_{k'=1}^k \hat{r}_{k'}$. Note that when $\hat{r}_k$ is fixed, the minimization of $\sum_{k'=1}^{k-1} \hat{r}_{k'}$ is independent of $\hat{r}_{k+1}, ..., \hat{r}_K$. If $k < K$, the algorithm computes the minimum $\sum_{k'=k}^K \hat{r}_{k'}$ for each possible value of $\hat{r}_K$. Then the algorithm computes the minimal objective if a feasible solution exists given $\hat{r}_k = \frac{1}{m_{\text{ind}^{-1}(k)}}$. At the end of the iteration with k, the algorithm updates the minimal objective seen so far.

B.2. Proof of The Correctness of Algorithm 2

Proof of Theorem 4.4. Say $\boldsymbol{r}^*$ is a solution of the optimization problem in Algorithm 2. By Lemma 2 in Li et al. (2021), the schedule that Algorithm 2 outputs is an AUS cycle of length $\frac{\sum_{k=1}^K r_k^*}{r_K^*}$. Since $\boldsymbol{r}^*$ satisfies (7) and (8), the length can be upper bounded as $\frac{\sum_{k=1}^K r_k^*}{r_K^*} \leq \frac{1}{r_K^*} \leq m_{\text{ind}^{-1}(K)} = \max_{k \in [K]} m_k = ||\boldsymbol{m}||_\infty$. Consider a cyclic schedule composed of such AUS cycles. To show that this cyclic schedule is feasible for $\boldsymbol{m}$, it suffices to show that $\forall k \in [K], i \geq 1 : I_k^i \leq m_{\text{ind}^{-1}(k)}$. We prove this by contradiction. For any k, suppose $\exists i : I_k^i > m_{\text{ind}^{-1}(k)}$, we have that $\forall i \geq 1 : I_k^i \geq m_{\text{ind}^{-1}(k)}$ by the definition of AUS. We consider the first AUS cycle in the schedule. By Algorithm 1 in Li et al. (2021), in the AUS cycle, $n_k = \frac{r_k^*}{r_K^*}$ for any k and $n = \frac{\sum_{k=1}^K r_k^*}{r_K^*}$. The next cycle in the schedule is just a repetition of the current cycle. So $\forall i \geq 1 : I_k^i \geq m_{\text{ind}^{-1}(k)}$ implies $\forall i \in [n_k], a_{[\tau_k^i \mod n]+1}, ..., a_{[(\tau_k^i + m_{\text{ind}^{-1}(k)} - 2) \mod n]+1} \neq k$. The i-th pull of arm k is followed by at least $m_{\text{ind}^{-1}(k)} - 1$ pulls of other arms for $i < n_k$, while the n_k-th pull of arm k implies there are at least $m_{\text{ind}^{-1}(k)} - 1$ pulls of other arms in the set of time slots $[\tau_k^1] \cup \{\tau_k^{n_k} + 1, ..., n\}$. Thus we have $m_{\text{ind}^{-1}(k)} n_k < n$ since $\exists i : I_k^i > m_{\text{ind}^{-1}(k)}$. Let r_k denote the average rate of arm k's appearance in the schedule, then $r_k = \frac{n_k}{n} < \frac{1}{m_{\text{ind}^{-1}(k)}}$. Thus the rate r_k satisfies

$$\frac{1}{m_{\text{ind}^{-1}(k)}} > r_k = \frac{r_k^*/r_K^*}{\sum_{k=1}^K r_k^*/r_K^*} \geq r_k^* \geq \frac{1}{m_{\text{ind}^{-1}(k)}}.$$

A contradiction occurs. So we have $\forall k \in [K], i \geq 1 : I_k^i \leq m_{\text{ind}^{-1}(k)}$ and the AUS cycle is feasible given the patience vector $\boldsymbol{m}$. □

C. Proofs for The Hardness Result

Proof of Lemma 3.3. Let $\bar{t} = \lambda K$. Suppose that $Q_{(K,\boldsymbol{m})} > \bar{t} + 1$. We have that from $t = 1$ to $t = \bar{t} + 1$ no arm leaves. For each arm i, say t_i^k is the time that it is pulled the k-th time. Consider the sequence $t = 0, t_i^1, ..., t_i^{T_i(\bar{t})}, \bar{t} + 1$. Say j, j' are

Algorithm 3 A Dynamic Programming Solution to the Optimization Problem in Algorithm 2

```
Input: Patience vector $\boldsymbol{m}$, the index mapping $\text{ind}(\cdot)$
$\boldsymbol{r}^* \leftarrow (1, ..., 1), \text{min\_len}^* \leftarrow +\infty$
for $k = 1, 2, ..., K$ do
    $\hat{\boldsymbol{r}} \leftarrow (1, ..., 1), \text{min\_len} \leftarrow +\infty$
    $\hat{r}_k \leftarrow \frac{1}{m_{\text{ind}^{-1}(k)}}, \mathcal{D}_k \leftarrow \{\hat{r}_k\}, \text{DP}(\hat{r}_k, k) \leftarrow \frac{1}{m_{\text{ind}^{-1}(k)}}$
    if $k > 1$ then
        $\mathcal{D}_{k'} \leftarrow \{l\hat{r}_k | \frac{1}{m_{\text{ind}^{-1}(k')}} \leq l\hat{r}_k \leq 1, l \in \mathbb{N}^+\}$ for each $k' = 1, ..., k-1$
        for $k' = k-1, ..., 1$ do
            for $r \in \mathcal{D}_{k'}$ do
                $\text{DP}(r, k') \leftarrow r + \min_{r': r' \in \mathcal{D}_{k'+1}, \frac{r}{r'} \in \mathbb{N}^+} \text{DP}(r', k'+1)$
                $\text{BWD}(r, k') \leftarrow \arg\min_{r': r' \in \mathcal{D}_{k'+1}, \frac{r}{r'} \in \mathbb{N}^+} \text{DP}(r', k'+1)$
            end for
        end for
        if $\min_{r \in \mathcal{D}_1} \text{DP}(r, 1) > 1$ then
            Continue
        end if
        $\hat{r}_1 \leftarrow \arg\min_{r \in \mathcal{D}_1} \text{DP}(r, 1)$
        for $k' = 2, ..., k-1$ do
            $\hat{r}_{k'} \leftarrow \text{BWD}(\hat{r}_{k'-1}, k'-1)$
        end for
    end if
    if $k < K$ then
        $\mathcal{D}_{k'} \leftarrow \{\frac{\hat{r}_k}{l} | \frac{1}{m_{\text{ind}^{-1}(k')}} \leq \frac{\hat{r}_k}{l}, l \in \mathbb{N}^+\}$ for each $k' = k+1, ..., K$
        for $k' = k+1, ..., K$ do
            for $r \in \mathcal{D}_{k'}$ do
                $\text{DP}(r, k') \leftarrow r + \min_{r': r' \in \mathcal{D}_{k'-1}, \frac{r'}{r} \in \mathbb{N}^+} \text{DP}(r', k'-1)$
                $\text{BWD}(r, k') \leftarrow \arg\min_{r': r' \in \mathcal{D}_{k'-1}, \frac{r'}{r} \in \mathbb{N}^+} \text{DP}(r', k'-1)$
            end for
        end for
    end if
    if $\min_{r \in \mathcal{D}_K} \sum_{k'=1}^{k-1} \hat{r}_k + \text{DP}(r, K) \leq 1$ then
        $\text{min\_len} \leftarrow \min_{r \in \mathcal{D}_K : \sum_{k'=1}^{k-1} \hat{r}_k + \text{DP}(r,K) \leq 1} r^{-1}[\sum_{k'=1}^{k-1} \hat{r}_k + \text{DP}(r, K)]$
    end if
    if $\text{min\_len} < \text{min\_len}^*$ then
        $\text{min\_len}^* \leftarrow \text{min\_len}$
        $r^*_{k'} \leftarrow \hat{r}_{k'}$ for each $k' = 1, ..., k$
        $r^*_K \leftarrow \arg\min_{r \in \mathcal{D}_K : \sum_{k'=1}^{k-1} \hat{r}_k + \text{DP}(r,K) \leq 1} r^{-1}[\sum_{k'=1}^{k-1} \hat{r}_k + \text{DP}(r, K)]$
        for $k' = K-1, ..., k+1$ do
            $r^*_{k'} \leftarrow \text{BWD}(r^*_{k'+1}, k'+1)$
        end for
    end if
end for
Return $\boldsymbol{r}^*$
```

consecutive items in the sequence, then we have $j' - j \leq m_i$. This implies that for $\forall i \in [K]$,

$$T_i(\bar{t}) \geq \left\lfloor \frac{\bar{t}}{m_i} \right\rfloor > \frac{\bar{t}}{m_i} - 1$$

$$\frac{T_i(\bar{t}) + 1}{\bar{t}} > \frac{1}{m_i}.$$

By definition $\bar{t} = \sum_{i=1}^{K} T_i(\bar{t})$. Summing over i, we have

$$1 + \frac{1}{\lambda} = 1 + \frac{K}{\lambda K} = \frac{K + \sum_{i=1}^{K} T_i(\bar{t})}{\bar{t}} >$$
$$\sum_{i=1}^{K} \frac{1}{m_i} = l(K, \boldsymbol{m}) > 1 + \frac{1}{\lambda}.$$

A contradiction occurs. Thus we can conclude that $Q_{(K,\boldsymbol{m})} \leq \bar{t} + 1 = \lambda K + 1$.

□

Proof of Theorem 3.2. Given an unschedulable configuration $(K, \boldsymbol{m})$, i.e. $Q_{(K,\boldsymbol{m})} \leq T$, after any bandit algorithm $A \in \mathcal{A}$ pulls $Q_{(K,\boldsymbol{m})}$ times, there must be at least 1 arm that has left. We mainly focus on these first $Q_{(K,\boldsymbol{m})}$ pulls. Define a measurable space $(\Omega_{Q_{(K,\boldsymbol{m})}}, \mathcal{F}_{Q_{(K,\boldsymbol{m})}})$ where $\Omega_{Q_{(K,\boldsymbol{m})}} = ([K] \times \mathbb{R})^{Q_{(K,\boldsymbol{m})}} \subset \mathbb{R}^{2Q_{(K,\boldsymbol{m})}}$ and $\mathcal{F}_{Q_{(K,\boldsymbol{m})}}$ is the σ-algebra of the subsets of $\Omega_{Q_{(K,\boldsymbol{m})}}$. Fix a positive constant $\Delta \leq \bar{\Delta}$, we construct $\boldsymbol{\mu}_i \in \Xi$, $i \in [K]$ such that

$$\boldsymbol{\mu}_i = (0, 0, ..., \underbrace{\Delta}_{\text{the } i^{\text{th}} \text{ entry}}, 0, ..., 0).$$

Say we run any A on some $\boldsymbol{\mu}_i$. Define Ψ as the set of all measurable mappings ψ that maps from $\Omega_{Q_{(K,\boldsymbol{m})}}$ to $[K]$. We want to find a lower bound on

$$\inf_{\psi \in \Psi} \max_{j \in [K]} \Pr_{A,(K,\boldsymbol{m},\boldsymbol{\mu}_j)} (\psi = j),$$

where $\Pr_{A,(K,\boldsymbol{m},\boldsymbol{\mu}_i)}, \forall i \in [K]$ are probability measures over $(\Omega_{Q_{(K,\boldsymbol{m})}}, \mathcal{F}_{Q_{(K,\boldsymbol{m})}})$. Our derivation is similar with that of Fano's inequality in Rigollet & Hütter (2019). We write $\Pr_{A,(K,\boldsymbol{m},\boldsymbol{\mu}_j)}$ as P_j for notational simplicity. Fix $\psi \in \Psi$, define

$$p_j = P_j(\psi \neq j), \quad \bar{p} = \frac{1}{K} \sum_{j=1}^{K} p_j \in [0, 1], \quad q_j = \frac{1}{K} \sum_{k=1}^{K} P_k(\psi \neq j),$$

$$\bar{q} = \frac{1}{K} \sum_{j=1}^{K} q_j = \frac{1}{K^2} \sum_{j,k=1}^{K} P_k(\psi \neq j) = \frac{1}{K^2} \sum_{j,k=1}^{K} 1 - P_k(\psi = j)$$
$$= 1 - \frac{K}{K^2} = \frac{K-1}{K}.$$

Let $kl(p, p')$ denote KL divergence between Bernoulli distributions with mean p, p', respectively. We have

$$kl(p, p') = p \ln \frac{p}{p'} + (1 - p) \ln \frac{1 - p}{1 - p'}.$$

Denote the entropy of Bernoulli distribution $h(p) = p \ln p + (1 - p) \ln(1 - p)$. We have on the one hand,

$$\begin{aligned} kl(\bar{p}, \bar{q}) =& h(\bar{p}) - \bar{p} \ln \bar{q} - (1 - \bar{p}) \ln(1 - \bar{q}) \\ \geq& h(\bar{p}) - \bar{p} \ln \bar{q} = h(\bar{p}) + \bar{p} \ln \frac{K}{K-1} =: f_K(\bar{p}). \end{aligned}$$

On the other hand, we derive an upper bound for $kl(\bar{p}, \bar{q})$, by the convexity of KL divergence,

$$kl(\bar{p}, \bar{q}) \leq \frac{1}{K}\sum_{j=1}^{K} kl(p_j, q_j) \leq \frac{1}{K^2}\sum_{j,k=1}^{K} kl(P_j(\psi \neq j), P_k(\psi \neq j)).$$

Fix any j, k, we derive upper bound for $kl(P_j(\psi \neq j), P_k(\psi \neq j))$. We observe that probability measures P_j, P_k over $(\Omega_{Q_{(K,\boldsymbol{m})}}, \mathcal{F}_{Q_{(K,\boldsymbol{m})}})$ induce probability measures $\text{Ber}(P_j(\{\omega \in \Omega_{Q_{(K,\boldsymbol{m})}} : \psi(\omega) \neq j\})), \text{Ber}(P_k(\{\omega \in \Omega_{Q_{(K,\boldsymbol{m})}} : \psi(\omega) \neq j\}))$ over measurable space $(\{0,1\}, 2^{\{0,1\}})$. Then by data processing inequality in Lou & Goldfeld (2020), especially their Example 7.1.2(i), we have

$$\begin{aligned}\text{KL}(P_j, P_k) &\geq \text{KL}(\text{Ber}(P_j(\psi \neq j)), \text{Ber}(P_k(\psi \neq j))) \\ &= kl(P_j(\psi \neq j), P_k(\psi \neq j)).\end{aligned}$$

Some sequences of $a_1, a_2, ..., a_{Q_{(K,\boldsymbol{m})}}$ are impossible due to the leaving behaviour of arms, depending only on $(K, \boldsymbol{m})$. However, an arm sequence is possible with respect to $(K, \boldsymbol{m}, \boldsymbol{\mu}_j)$ if and only if it is possible with respect to $(K, \boldsymbol{m}, \boldsymbol{\mu}_k)$. As a result, P_j is absolutely continuous with respect to P_k. With this fact in mind, we can use Lemma 15.1 in Lattimore & Szepesvári (2020) to show that

$$\begin{aligned}\text{KL}(P_j, P_k) &= \sum_{i=1}^{K} \mathbb{E}_j[T_i(Q_{(K,\boldsymbol{m})})]\text{KL}(\mathcal{N}(\boldsymbol{\mu}_{j,i}, 1), \mathcal{N}(\boldsymbol{\mu}_{k,i}, 1)) \\ &= \frac{\Delta^2}{2}(\mathbb{E}_j[T_j(Q_{(K,\boldsymbol{m})})] + \mathbb{E}_j[T_k(Q_{(K,\boldsymbol{m})})]) \\ &\leq \frac{\Delta^2}{2} Q_{(K,\boldsymbol{m})}.\end{aligned}$$

Merging the derivations above, we obtain that

$$f_K(\bar{p}) \leq \frac{1}{K^2}\sum_{j,k=1}^{K} \frac{\Delta^2}{2} Q_{(K,\boldsymbol{m})} = \frac{\Delta^2}{2} Q_{(K,\boldsymbol{m})}.$$

Now we study the function $f_K(p)$ to obtain an upper bound for $\bar{p}$.

$$\begin{aligned}f_K(p) &= p \ln \frac{K}{K-1} + p\ln p + (1-p)\ln(1-p), \\ f_K'(p) &= \ln \frac{K}{K-1} + \ln p + 1 - \ln(1-p) - 1 \\ &= \ln \frac{K}{K-1} + \ln \frac{p}{1-p}.\end{aligned}$$

We see that $f_K(0) = 0$, $f_K(1) = \ln \frac{K}{K-1}$, f_K is monotonically decreasing in $[0, \frac{K-1}{2K-1}]$ and monotonically increasing in $[\frac{K-1}{2K-1}, 1]$. Thus $\forall c \in (0, \ln \frac{K}{K-1}]$, there is a unique $p \in (\frac{K-1}{2K-1}, 1]$ such that $f_K(p) = c$. Setting $\Delta = \sqrt{\frac{1}{Q_{(K,\boldsymbol{m})}} \ln \frac{K}{K-1}}$, let $f_K^{-1}(\frac{1}{2}\ln\frac{K}{K-1})$ denote the unique p such that $f_K(p) = \frac{Q_{(K,\boldsymbol{m})}}{2}\frac{1}{Q_{(K,\boldsymbol{m})}}\ln\frac{K}{K-1} = \frac{1}{2}\ln\frac{K}{K-1}$. We obtain that

$$\frac{1}{K}\sum_{j=1}^{K} P_j(\psi \neq j) = \bar{p} \leq f_K^{-1}(\frac{1}{2}\ln\frac{K}{K-1}).$$

Thus

$$\inf_{\psi\in\Psi}\max_{j\in[K]} P_j(\psi = j) \geq \inf_{\psi\in\Psi}\frac{1}{K}\sum_{j=1}^{K} P_j(\psi = j)$$

$$= \inf_{\psi\in\Psi}\frac{1}{K}\sum_{j=1}^{K} 1 - P_j(\psi \neq j)$$

$$\geq 1 - f_K^{-1}(\frac{1}{2}\ln\frac{K}{K-1}).$$

Define ψ_L as the measurable map that returns the first leaving arm during the first $Q_{(K,\boldsymbol{m})}$ pulls, breaking tie by returning the first leaving arm with the smallest arm index. We claim that, for any bandit policy A and configuration $(K, \boldsymbol{m})$, there is $j \in [K]$ such that $P_j(\psi_L = j) \geq 1 - f_K^{-1}(\frac{1}{2}\ln\frac{K}{K-1})$. Suppose not, then there exist A, $(K, \boldsymbol{m})$ that for all $j \in [K]$, $P_j(\psi_L = j) < 1 - f_K^{-1}(\frac{1}{2}\ln\frac{K}{K-1})$, then setting $\Delta = \sqrt{\frac{1}{Q_{(K,\boldsymbol{m})}}\ln\frac{K}{K-1}}$,

$$
\begin{aligned}
1 - f_K^{-1}(\frac{1}{2}\ln\frac{K}{K-1}) &> \max_{j\in[K]} P_j(\psi_L = j) \\
\geq \inf_{\psi\in\Psi}\max_{j\in[K]} P_j(\psi = j) & \\
\geq 1 - f_K^{-1}(\frac{1}{2}\ln\frac{K}{K-1}). &
\end{aligned}
$$

A contradiction occurs, so we proved the claim. For any algorithm A, say arm i_A satisfies $P_{i_A}(\psi_L = i_A) \geq 1 - f_K^{-1}(\frac{1}{2}\ln\frac{K}{K-1})$. Now we derive the lower bound for expected regret when the ground truth reward mean vector is $\boldsymbol{\mu}_{i_A}$

$$
\begin{aligned}
R_T(A,(K,\boldsymbol{m},\boldsymbol{\mu}_{i_A})) &= \mathbb{E}_{i_A}\Big[\sum_{t=1}^{T}\Delta_{a_t}\Big] = \mathbb{E}_{i_A}\Big[\sum_{k\neq i_A}^{K}\Delta T_k(T)\Big] \\
&= \sqrt{\frac{1}{Q_{(K,\boldsymbol{m})}}\ln\frac{K}{K-1}}\sum_{k\neq i_A}^{K}\mathbb{E}_{i_A}\Big[T_k(T)\Big] \\
&= \sqrt{\frac{1}{Q_{(K,\boldsymbol{m})}}\ln\frac{K}{K-1}}\mathbb{E}_{i_A}\Big[T - T_{i_A}(T)\Big] \\
&= \sqrt{\frac{1}{Q_{(K,\boldsymbol{m})}}\ln\frac{K}{K-1}}\mathbb{E}_{i_A}\Big[T - T_{i_A}(T)|\psi_L = i_A\Big]P_{i_A}(\psi_L = i_A) \\
&+ \sqrt{\frac{1}{Q_{(K,\boldsymbol{m})}}\ln\frac{K}{K-1}}\mathbb{E}_{i_A}\Big[T - T_{i_A}(T)|\psi_L \neq i_A\Big]P_{i_A}(\psi_L \neq i_A) \\
&\geq \sqrt{\frac{1}{Q_{(K,\boldsymbol{m})}}\ln\frac{K}{K-1}}\Big[1 - f_K^{-1}(\frac{1}{2}\ln\frac{K}{K-1})\Big](T - Q_{(K,\boldsymbol{m})}),
\end{aligned}
$$

where $\mathbb{E}_j$ is the expectation when $A, (K, \boldsymbol{m}, \boldsymbol{\mu}_j)$ is given. Finally, we note that the bound above holds for any A, and

$$
R_T(A,(K,\boldsymbol{m},\boldsymbol{\mu}_{i_A})) \leq \sup_{\boldsymbol{\mu}\in\Xi} R_T(A,(K,\boldsymbol{m},\boldsymbol{\mu})).
$$

So we can make our conclusion, for any unschedulable profile $(K, \boldsymbol{m})$, the minimax lower bound for the expected regret is

$$
\min_{A\in\mathcal{A}}\sup_{\boldsymbol{\mu}\in\Xi} R_T(A,(K,\boldsymbol{m},\boldsymbol{\mu})) \geq \sqrt{\frac{1}{Q_{(K,\boldsymbol{m})}}\ln\frac{K}{K-1}}\Big[1 - f_K^{-1}(\frac{1}{2}\ln\frac{K}{K-1})\Big](T - Q_{(K,\boldsymbol{m})}).
$$

□

D. Results for UCB Algorithm

In the standard regret analysis of UCB in the stochastic MAB setting (Lattimore & Szepesvári, 2020), a critical observation is that under a high probability event (usually called a ”Good” event) the number of pulls of sub-optimal arms is bounded. In our setup characterized by arm patience, arm k exits if there have been too many pulls of other arms after its last pull. However, if the optimal arm k^*'s patience m_{k^*} is even greater than the highest possible total number of sub-optimal pulls under the good event, then it is unlikely that arm k^* will exit under the good event. Thus, if arm k^* has good patience, the performance of UCB algorithm can be intuitively guaranteed. Inspired by this observation, we present the following expected regret upper bound for UCB algorithm in our setup.

Theorem D.1. *Run UCB algorithm with confidence parameter $\delta = 1/T^2$ in the Multi-Armed Bandit with Impatient Arms setup. If $m_{k^*} > \sum_{k\neq k^*}^{K} \lceil \frac{16 \ln T}{\Delta_k^2} \rceil$ the expected regret of UCB algorithm is upper bounded by*

$$R_T \leq K\Delta_{\max} + 3\sum_{k\neq k^*}^{K} \Delta_k + \sum_{k\neq k^*}^{K} \frac{16 \ln T}{\Delta_k},$$

where $\Delta_{max} := \max_{k\neq k^*} \Delta_k$.

Proof of Theorem D.1. Define $\mathcal{D}$ as the event that arm k^* is ignored for at least m_{k^*} times, formally

$$\mathcal{D} = \Big\{\exists t \in \{1, 2, ..., T + 1 - m_{k^*}\} : a_t, a_{t+1}, ..., a_{t+m_{k^*}-1} \neq k^*\Big\}.$$

The expected regret is decomposed with respect to event $\mathcal{D}$,

$$\begin{aligned} R_T =&\mathbb{E}\big[\sum_{t=1}^{T} \mu^* - \sum_{t=1}^{T} \mu_{a_t}\big] \\ =&\mathbb{E}\big[\sum_{t=1}^{T} \Delta_{a_t}\big] \\ =&\mathbb{E}\big[\sum_{t=1}^{T} \Delta_{a_t}\mathbb{I}\{\bar{\mathcal{D}}\}\big] + \mathbb{E}\big[\sum_{t=1}^{T} \Delta_{a_t}\mathbb{I}\{\mathcal{D}\}\big] \\ \leq&\Delta_{\max} T \Pr(\mathcal{D}) + R_T^{\text{Base}}, \end{aligned}$$

where R_T^{Base} denotes $\mathbb{E}\big[\sum_{t=1}^{T} \Delta_{a_t}\mathbb{I}\{\bar{\mathcal{D}}\}\big]$, whose analysis resembles that of UCB algorithm in the vanilla MAB setting. Following the standard analysis of UCB algorithm, we can define ”good” events when the sample means of each arm lie in their corresponding confidence intervals. Now we define ’good’ events.

$$\mathcal{E}_{k^*} := \Big\{\mu^* < \min_{n\in[T]} \hat{\mu}_{k^*,n} + \sqrt{\frac{2\ln 1/\delta}{n}}\Big\},$$

$$\mathcal{E}_{k} := \Big\{\hat{\mu}_{k,u_k} + \sqrt{\frac{2\ln 1/\delta}{u_k}} < \mu^*\Big\}, \quad \forall k \neq k^*.$$

If m_{k^*} is large enough, formally $m_{k^*} > \sum_{k\neq k^*}^{K} u_k$, by total probability rule,

$$\begin{aligned} \Pr(\mathcal{D}) =& \Pr\Big(\mathcal{D}\Big|\bigcap_{k=1}^{K} \mathcal{E}_k\Big) \Pr\Big(\bigcap_{k=1}^{K} \mathcal{E}_k\Big) + \Pr\Big(\mathcal{D}\Big|\bigcup_{k=1}^{K} \bar{\mathcal{E}}_k\Big) \Pr\Big(\bigcup_{k=1}^{K} \bar{\mathcal{E}}_k\Big) \\ \leq& \Pr\Big(\mathcal{D}\Big|\bigcap_{k=1}^{K} \mathcal{E}_k\Big) + \Pr\Big(\bigcup_{k=1}^{K} \bar{\mathcal{E}}_k\Big) \\ \leq& \Pr\Big(\mathcal{D}\Big|\bigcap_{k=1}^{K} \mathcal{E}_k\Big) + \sum_{k=1}^{K} \Pr\Big(\bar{\mathcal{E}}_k\Big). \end{aligned}$$

Conditioned on event $\bigcap_{k=1}^{K} \mathcal{E}_k$, if $\exists t\ s.t.\ a_t, a_{t+1}, ..., a_{t+m_{k^*}-1} \neq k^*$, there must be at least one arm denoted k, that it is selected more than u_k times even during the time interval from t to $t + m_{k^*} - 1$. However under event $\mathcal{E}_{k^*}$, $\mathcal{E}_k$, and the fact that arm k may deviate in advance, arm k cannot be selected for more than u_k times. There is a contradiction, so $\Pr(\mathcal{D}| \bigcap_{k=1}^{K} \mathcal{E}_k) = 0$. We see that $\Pr(\mathcal{D}) \leq \sum_{k=1}^{K} \Pr(\bar{\mathcal{E}}_k)$ as a result.

On the other hand,

$$
\begin{aligned}
R_T^{\text{Base}} =&\mathbb{E}\big[\sum_{t=1}^{T}\Delta_{a_t}\mathbb{I}\{\bar{\mathcal{D}}\}\big]\\
=&\mathbb{E}\big[\sum_{k\neq k^*}^{K}\sum_{t=1}^{T}\Delta_k\mathbb{I}\{a_t=k\}\mathbb{I}\{\bar{\mathcal{D}}\}\big]\\
=&\sum_{k\neq k^*}^{K}\Delta_k\mathbb{E}\big[T_k(T)\mathbb{I}\{\bar{\mathcal{D}}\}\big]\\
=&\sum_{k\neq k^*}^{K}\Delta_k\Big(\mathbb{E}\big[T_k(T)\mathbb{I}\{\mathcal{E}_{k^*},\mathcal{E}_k\}\mathbb{I}\{\bar{\mathcal{D}}\}\big]+\mathbb{E}\big[T_k(T)\mathbb{I}\{\bar{\mathcal{E}}_{k^*}\cup\bar{\mathcal{E}}_k\}\mathbb{I}\{\bar{\mathcal{D}}\}\big]\Big)\\
\leq&\sum_{k\neq k^*}^{K}\Delta_k\Big(u_k+\mathbb{E}\big[T_k(T)\mathbb{I}\{\bar{\mathcal{E}}_{k^*}\cup\bar{\mathcal{E}}_k\}\mathbb{I}\{\bar{\mathcal{D}}\}\big]\Big)\\
\leq&\sum_{k\neq k^*}^{K}\Delta_k\Big(u_k+T\mathbb{E}\big[\mathbb{I}\{\bar{\mathcal{E}}_{k^*}\cup\bar{\mathcal{E}}_k\}\mathbb{I}\{\bar{\mathcal{D}}\}\big]\Big)\\
\leq&\sum_{k\neq k^*}^{K}\Delta_k\Big(u_k+T\Pr(\bar{\mathcal{E}}_{k^*}\cup\bar{\mathcal{E}}_k)\Big).
\end{aligned}
$$

The first inequality holds since under event $\bar{\mathcal{D}}$, arm k^* never leaves the game and under event $\mathcal{E}_{k^*}\cap\mathcal{E}_k$, no matter whether arm k derivates, it can be selected for no more than u_k times. Even if it is pulled for u_k times, it will never be pulled again since arm k^*'s UCB index is always larger. By Chernoff's inequality, say $c_k, \forall k\neq k^*$ satisfy $\Delta_k-\sqrt{\frac{2\ln 1/\delta}{u_k}}\geq c_k\Delta_k$,

$$
\begin{aligned}
\Pr(\bar{\mathcal{E}}_{k^*}) =&\Pr\Big(\bigcup_{n\in[T]}\{\mu^*\geq\hat{\mu}_{k^*,n}+\sqrt{\frac{2\ln 1/\delta}{n}}\}\Big)\leq\sum_{n=1}^{T}\Pr\Big(\mu^*\geq\hat{\mu}_{k^*,n}+\sqrt{\frac{2\ln 1/\delta}{n}}\Big)\leq T\delta,\\
\Pr(\bar{\mathcal{E}}_k)\leq&\Pr(\hat{\mu}_{k,u_k}-\mu_k\geq c_k\Delta_k)\leq\exp\Big(-\frac{u_kc_k^2\Delta_k^2}{2}\Big),\ \ \forall k\neq k^*.
\end{aligned}
$$

Then for the expected regret when $m_{k^*}>\sum_{k\neq k^*}^{K}u_k$,

$$
\begin{aligned}
R_T\leq&\sum_{k\neq k^*}^{K}\Delta_k\Big(u_k+T\Pr(\bar{\mathcal{E}}_{k^*}\cup\bar{\mathcal{E}}_k)\Big)+\Delta_{\max}T\sum_{k=1}^{K}\Pr(\bar{\mathcal{E}}_k)\\
\leq&\sum_{k\neq k^*}^{K}\Delta_ku_k+\sum_{k\neq k^*}^{K}\Delta_kT\big(T\delta+\exp\Big(-\frac{u_kc_k^2\Delta_k^2}{2}\Big)\big)+\Delta_{\max}TT\delta+\Delta_{\max}T\sum_{k\neq k^*}^{K}\exp\Big(-\frac{u_kc_k^2\Delta_k^2}{2}\Big)\\
=&\sum_{k\neq k^*}^{K}\Delta_ku_k+\sum_{k\neq k^*}^{K}\Delta_k+\Delta_{\max}+\sum_{k\neq k^*}^{K}(\Delta_k+\Delta_{\max})T\exp\Big(-\frac{u_kc_k^2\Delta_k^2}{2}\Big).
\end{aligned}
$$

We set $u_k=\lceil\frac{2\ln 1/\delta}{(1-c_k)^2\Delta_k^2}\rceil$ and obtain

$$
R_T\leq\sum_{k\neq k^*}^{K}\Delta_k\Big\lceil\frac{2\ln 1/\delta}{(1-c_k)^2\Delta_k^2}\Big\rceil+\sum_{k\neq k^*}^{K}\Delta_k+\Delta_{\max}+\sum_{k\neq k^*}^{K}(\Delta_k+\Delta_{\max})T^{1-\frac{2c_k^2}{(1-c_k)^2}}.
$$

By setting $c_k = \frac{1}{2}, \forall k \neq k^*$, we obtain

$$\begin{aligned} R_T \leq& \sum_{k \neq k^*}^{K} \Delta_k \left\lceil \frac{8 \ln 1/\delta}{\Delta_k^2} \right\rceil + \sum_{k \neq k^*}^{K} \Delta_k + \Delta_{\max} + \sum_{k \neq k^*}^{K} (\Delta_k + \Delta_{\max}) \\ \leq& K\Delta_{\max} + 3\sum_{k \neq k^*}^{K} \Delta_k + \sum_{k \neq k^*}^{K} \frac{16 \ln T}{\Delta_k}. \end{aligned}$$

□

In Theorem D.1, we derive a problem-dependent upper bound for a fixed $\boldsymbol{\mu}$. Based on Theorem D.1, we further present a problem-independent expected regret upper bound for a general set of mean reward vectors in Corollary D.2.

Corollary D.2. *Run UCB algorithm with confidence parameter $\delta = 1/T^2$ in the Multi-Armed Bandit with Impatient Arms setup. For a constant $\epsilon > 0$, consider a set of mean reward vectors $\Xi_\epsilon = \left\{\boldsymbol{\mu} \,\middle|\, \epsilon \leq \Delta_k \leq \bar{\Delta}, \forall k \neq k^*\right\}$. For any instance $(K, \boldsymbol{m}, \boldsymbol{\mu})$ satisfying: 1. $\boldsymbol{\mu} \in \Xi_\epsilon$, 2. $m_k > (K-1)\lceil 16\epsilon^{-2} \ln T \rceil, \forall k \in [K]$, the expected regret of UCB algorithm is uniformly upper bounded by*

$$R_T \leq K\Delta_{\max} + 3\sum_{k \neq k^*}^{K} \Delta_k + 8\sqrt{(K-1)T \ln T},$$

where $\Delta_{max} := \max_{k \neq k^*} \Delta_k$.

In Corollary D.2, we show that UCB algorithm maintains almost the same performance as in stochastic MAB setting, under the assumption that all arm have relatively high level of patience.

Proof of Corollary D.2. $m_k > (K-1)\lceil 16\epsilon^{-2} \ln T \rceil, \forall k \in [K]$ implies that $\forall \boldsymbol{\mu} \in \Xi_\epsilon$, we have $m_{k^*} > (K-1)\lceil \frac{16 \ln T}{\epsilon^2} \rceil \geq \sum_{k \neq k^*}^{K} \lceil \frac{16 \ln T}{\Delta_k^2} \rceil$. We consider two cases for the value of ϵ:

(1) $\epsilon \geq \Delta := \sqrt{16(K-1)T^{-1} \ln T}$. In this case, $\boldsymbol{\mu}$ satisfies that $\Delta_k \geq \sqrt{16(K-1)T^{-1} \ln T}, \forall k \neq k^*$. We directly adopt Theorem D.1 and obtain

$$\begin{aligned} R_T \leq& K\Delta_{\max} + 3\sum_{k \neq k^*}^{K} \Delta_k + \sum_{k \neq k^*}^{K} \frac{16 \ln T}{\Delta_k} \\ \leq& K\Delta_{\max} + 3\sum_{k \neq k^*}^{K} \Delta_k + 4\sqrt{(K-1)T \ln T} \\ \leq& 8\sqrt{(K-1)T \ln T} + 3\sum_{k \neq k^*}^{K} \Delta_k + K\Delta_{\max}. \end{aligned}$$

(2) $\epsilon < \Delta := \sqrt{16(K-1)T^{-1} \ln T}$. Following the derivation in the proof of Theorem D.1, the operation on the first term

of the regret decomposition remains the same, since $m_{k^*} > \sum_{k\neq k^*}^{K} \lceil \frac{16\ln T}{\Delta_k^2} \rceil$. So we focus on the second term,

$$\begin{aligned}
R_T^{\text{Base}} = &\sum_{k\neq k^*}^{K} \Delta_k \mathbb{E}\big[T_k(T)\mathbb{I}\{\bar{\mathcal{D}}\}\big] = \sum_{k\neq k^*:\Delta_k\geq\Delta} \Delta_k \mathbb{E}\big[T_k(T)\mathbb{I}\{\bar{\mathcal{D}}\}\big] + \sum_{k\neq k^*:\Delta_k<\Delta} \Delta_k \mathbb{E}\big[T_k(T)\mathbb{I}\{\bar{\mathcal{D}}\}\big] \\
&<\Delta T + \sum_{k\neq k^*:\Delta_k\geq\Delta} \Delta_k \Big(\mathbb{E}\big[T_k(T)\mathbb{I}\{\mathcal{E}_{k^*}, \mathcal{E}_k\}\mathbb{I}\{\bar{\mathcal{D}}\}\big] + \mathbb{E}\big[T_k(T)\mathbb{I}\{\bar{\mathcal{E}}_{k^*} \cup \bar{\mathcal{E}}_k\}\mathbb{I}\{\bar{\mathcal{D}}\}\big]\Big) \\
&\leq\Delta T + \sum_{k\neq k^*:\Delta_k\geq\Delta} \Delta_k \Big(u_k + \mathbb{E}\big[T_k(T)\mathbb{I}\{\bar{\mathcal{E}}_{k^*} \cup \bar{\mathcal{E}}_k\}\mathbb{I}\{\bar{\mathcal{D}}\}\big]\Big) \\
&\leq\Delta T + \sum_{k\neq k^*:\Delta_k\geq\Delta} \Delta_k \Big(u_k + T\mathbb{E}\big[\mathbb{I}\{\bar{\mathcal{E}}_{k^*} \cup \bar{\mathcal{E}}_k\}\mathbb{I}\{\bar{\mathcal{D}}\}\big]\Big) \\
&\leq\Delta T + \sum_{k\neq k^*:\Delta_k\geq\Delta} \Delta_k \Big(u_k + T\Pr(\bar{\mathcal{E}}_{k^*} \cup \bar{\mathcal{E}}_k)\Big).
\end{aligned}$$

Merging the two terms, we also have

$$\begin{aligned}
R_T \leq&\Delta T + \sum_{k\neq k^*:\Delta_k\geq\Delta} \Delta_k \Big(u_k + T\Pr(\bar{\mathcal{E}}_{k^*} \cup \bar{\mathcal{E}}_k)\Big) + \Delta_{\max} T \sum_{k=1}^{K} \Pr(\bar{\mathcal{E}}_k) \\
\leq&\Delta T + \sum_{k\neq k^*:\Delta_k\geq\Delta} \Delta_k u_k + \sum_{k\neq k^*}^{K} \Delta_k + \Delta_{\max} + \sum_{k\neq k^*}^{K} (\Delta_k + \Delta_{\max}) T \exp(-\frac{u_k c_k^2 \Delta_k^2}{2}) \\
\leq&\Delta T + \sum_{k\neq k^*:\Delta_k\geq\Delta} \Delta_k \Big(\frac{2\ln 1/\delta}{(1-c_k)^2\Delta_k^2} + 1\Big) + 2\sum_{k\neq k^*}^{K} \Delta_k + K\Delta_{\max} \\
\leq&\Delta T + \sum_{k\neq k^*:\Delta_k\geq\Delta} \frac{16\ln T}{\Delta_k} + 3\sum_{k\neq k^*}^{K} \Delta_k + K\Delta_{\max} \\
\leq&\Delta T + \frac{16(K-1)\ln T}{\Delta} + 3\sum_{k\neq k^*}^{K} \Delta_k + K\Delta_{\max} \\
=&8\sqrt{(K-1)T\ln T} + 3\sum_{k\neq k^*}^{K} \Delta_k + K\Delta_{\max},
\end{aligned}$$

with $c_k = \frac{1}{2}$ for any $k \neq k^*$.

□

Proof of Theorem 3.5. Fix constants $a_\beta, b_\beta, \kappa_T, g_T < T$, and $x_1 > x_2 > \mu^*$. The precise values of these constants will be specified later. We define some events for arm k^*,

$$\mathcal{E}_{k^*} := \mathcal{E}_{k^*}(x_1, x_2) = \big\{ \min_{n\in[\kappa_T - 1]} \text{UCB}_{k^*}(n) > x_1, \text{UCB}_{k^*}(\kappa_T) < x_2 \big\},$$

and for arm β,

$$\mathcal{E}_\beta := \mathcal{E}_\beta(x_1, x_2) = \big\{ x_2 < \text{UCB}_\beta(a_\beta) < x_1, \min_{n\in[b_\beta]} \text{UCB}_\beta(n) > x_2 \big\}.$$

We will show later that under $\mathcal{E}_{k^*} \cap \mathcal{E}_\beta$ with our value assignments, the optimal arm k^* is only pulled at most κ_T times before it leaves. The number of sub-optimal pulls is at least $T - \kappa_T$ in total. Set

$$x_1 = \mu^* + \sqrt{\frac{2\ln 1/\delta}{\kappa_T - 1}} > x_2 = \mu^* + \sqrt{\frac{2\ln 1/\delta}{\kappa_T}} - \frac{g_T}{\kappa_T} > \mu^*,$$

κ_T, g_T are positive and non-decreasing w.r.t T. Now we compute lower bounds for both $\Pr(\mathcal{E}_{k^*}(x_1,x_2))$ and $\Pr(\mathcal{E}_\beta(x_1,x_2))$.

$$\begin{aligned}
&\Pr(\mathcal{E}_{k^*}(x_1,x_2)) = \Pr\Big(\min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1, \text{UCB}_{k^*}(\kappa_T) < x_2\Big)\\
=&\Pr\Big(\text{UCB}_{k^*}(\kappa_T) < x_2 \Big| \min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1\Big)\Pr\Big(\min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1\Big)\\
=&\Pr\Big(\sum_{j=1}^{\kappa_T} X_{k^*,j} - \mu^* < -g_T \Big| \min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1\Big)\Pr\Big(\min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1\Big)\\
=&\Pr\Big(X_{k^*,\kappa_T} - \mu^* + \sum_{j=1}^{\kappa_T-1} X_{k^*,j} - \mu^* < -g_T \Big| \min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1\Big)\Pr\Big(\min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1\Big)\\
=&\Pr\Big(\min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1\Big)\int_0^{+\infty}\Pr\Big(X_{k^*,\kappa_T} - \mu^* < -s - g_T \Big| S^{(k^*)}_{\kappa_T-1} = s, \min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1\Big)\\
&f\Big(S^{(k^*)}_{\kappa_T-1} = s \Big| \min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1\Big)ds,
\end{aligned}$$

where we denote that $S^{(i)}_t = \sum_{j=1}^t (X_{i,j} - \mu_i)$. Say $\Phi(\cdot)$ is the cdf of the standard Gaussian distribution. Since $X_{k^*,\kappa_T} - \mu^*$ is standard Gaussian and it is independent of any other random variables,

$$\Pr\Big(X_{k^*,\kappa_T} - \mu^* < -s - g_T \Big| S^{(k^*)}_{\kappa_T-1} = s, \min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1\Big) = \Phi(-s-g_T) = 1 - \Phi(s+g_T).$$

Now we consider the conditional probably density $f\Big(S^{(k^*)}_{\kappa_T-1} = s \Big| \min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1\Big)$. By Bayes' Theorem,

$$f\Big(S^{(k^*)}_{\kappa_T-1} = s \Big| \min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1\Big) = \frac{\Pr\Big(\min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1 \Big| S^{(k^*)}_{\kappa_T-1} = s\Big)}{\Pr\Big(\min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1\Big)} f(S^{(k^*)}_{\kappa_T-1} = s).$$

Since $S^{(k^*)}_{\kappa_T-1} \sim \mathcal{N}(0, \kappa_T - 1)$, we have $f(S^{(k^*)}_{\kappa_T-1} = s) = \frac{1}{\sqrt{2\pi(\kappa_T-1)}}\exp(-\frac{s^2}{2(\kappa_T-1)})$. It suffices to lower bounding $\Pr\Big(\min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1 \Big| S^{(k^*)}_{\kappa_T-1} = s\Big)$.

$$\begin{aligned}
&\Pr\Big(\min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1 \Big| S^{(k^*)}_{\kappa_T-1} = s\Big)\\
=&\Pr\Big(\forall n = 1,...,\kappa_T - 1 : S^{(k^*)}_n > (\frac{n}{\sqrt{\kappa_T - 1}} - \sqrt{n})\sqrt{2\ln 1/\delta} \Big| S^{(k^*)}_{\kappa_T-1} = s\Big)\\
\geq&\Pr\Big(\forall n = 1,...,\kappa_T - 1 : S^{(k^*)}_n > (1 - \frac{1}{\sqrt{\kappa_T - 1}})\frac{\sqrt{2\ln 1/\delta}}{\kappa_T - 2}(n - \kappa_T + 1) \Big| S^{(k^*)}_{\kappa_T-1} = s\Big)\\
=&1 - \Pr\Big(\exists n = 1,...,\kappa_T - 1 : S^{(k^*)}_n \leq (1 - \frac{1}{\sqrt{\kappa_T - 1}})\frac{\sqrt{2\ln 1/\delta}}{\kappa_T - 2}(n - \kappa_T + 1) \Big| S^{(k^*)}_{\kappa_T-1} = s\Big)\\
\geq&1 - \Pr\Big(\exists t \in [0, \kappa_T - 1] : B(t) \leq (1 - \frac{1}{\sqrt{\kappa_T - 1}})\frac{\sqrt{2\ln 1/\delta}}{\kappa_T - 2}(t - \kappa_T + 1) \Big| B(\kappa_T - 1) = s\Big),
\end{aligned}$$

where $\{B(t), t \geq 0\}$ is standard Brownian motion. The first inequality is by convexity. The last inequality is by the fact that the joint distribution of $S^{(k^*)}_1, ..., S^{(k^*)}_{\kappa_T-1}$ is identical to that of $B(1), ..., B(\kappa_T - 1)$. To proceed, we introduce a boundary-crossing property of Brownian bridge mentioned in Scheike (1992).

Lemma D.3 (Proposition 3, Scheike (1992))**.** *If $B(t)$ is standard Brownian motion, and $a > 0$, $T < \infty$, and $c < a + bT$, then*

$$\Pr\Big(\bigcup_{t\in[0,T]}(B(t) \geq a + bt)\Big| B(T) = c\Big) = \exp\Big(-2a(a + bT - c)/T\Big).$$

If $a \leq 0$, the probability is 1.

By setting

$$a = (\kappa_T - 1)(1 - \frac{1}{\sqrt{\kappa_T - 1}})\frac{\sqrt{2\ln 1/\delta}}{\kappa_T - 2}, \quad b = -(1 - \frac{1}{\sqrt{\kappa_T - 1}})\frac{\sqrt{2\ln 1/\delta}}{\kappa_T - 2}, \quad c = -s$$

and adopting Lemma D.3, we obtain

$$\Pr\Big(\min_{n\in[\kappa_T-1]} \text{UCB}_{k^*}(n) > x_1 \Big| S^{(k^*)}_{\kappa_T-1} = s\Big) \geq 1 - \exp\Big(- 2(1 - \frac{1}{\sqrt{\kappa_T - 1}})\frac{\sqrt{2\ln 1/\delta}}{\kappa_T - 2} s\Big).$$

Now we have

$$\begin{aligned}
\Pr(\mathcal{E}_{k^*}(x_1, x_2)) \geq \int_0^{+\infty} [1 - \Phi(s + g_T)] \Bigg[1 - \exp\Big(- 2(1 - \frac{1}{\sqrt{\kappa_T - 1}})\frac{\sqrt{2\ln 1/\delta}}{\kappa_T - 2} s\Big)\Bigg] f(S^{(k^*)}_{\kappa_T-1} = s) ds \\
\geq \Bigg[1 - \exp\Big(- 2(1 - \frac{1}{\sqrt{\kappa_T - 1}})\frac{\sqrt{2\ln 1/\delta}}{\kappa_T - 2} g_T\Big)\Bigg] \int_{g_T}^{+\infty} \frac{2}{\sqrt{4 + (s + g_T)^2} + s + g_T} \frac{1}{\sqrt{2\pi}} \exp\Big(- \frac{(s + g_T)^2}{2}\Big) \\
\frac{1}{\sqrt{2\pi(\kappa_T - 1)}} \exp\Big(- \frac{s^2}{2(\kappa_T - 1)}\Big) ds,
\end{aligned}$$

in the last inequality we use Komatu's lower bound mentioned in Duembgen (2010). To further derive a lower bound for $\Pr(\mathcal{E}_{k^*}(x_1, x_2))$, we consider

$$\begin{aligned}
&\int_{g_T}^{+\infty} \frac{2}{\sqrt{4 + (s + g_T)^2} + s + g_T} \frac{1}{\sqrt{2\pi}} \exp\Big(- \frac{(s + g_T)^2}{2}\Big) \exp\Big(- \frac{s^2}{2(\kappa_T - 1)}\Big) ds \\
\geq &\int_{g_T}^{+\infty} \frac{2s}{\sqrt{4 + 4s^2} + 2s} \frac{1}{\sqrt{2\pi}} \exp\Big(- \frac{(s + g_T)^2}{2}\Big) \exp\Big(- \frac{s^2}{2(\kappa_T - 1)} - \frac{s}{g_T}\Big) \frac{1}{g_T} ds \\
\geq &\frac{1}{\sqrt{1 + g_T^2} + g_T} \int_{g_T}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(- \frac{(s + g_T)^2}{2}\Big) \exp\Big(- \frac{s^2}{2(\kappa_T - 1)} - \frac{s}{g_T}\Big) ds \\
\geq &\frac{1}{\sqrt{1 + g_T^2} + g_T} \exp\left(\frac{1}{2}\frac{\kappa_T - 1}{\kappa_T}(g_T + \frac{1}{g_T})^2 - \frac{1}{2} g_T^2 \right) \\
&\times \int_{g_T}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(- \Big(\frac{1}{\sqrt{2}}\sqrt{\frac{\kappa_T}{\kappa_T - 1}} s + \frac{\sqrt{2}}{2}\sqrt{\frac{\kappa_T - 1}{\kappa_T}}(g_T + \frac{1}{g_T})\Big)^2 \right) ds \\
\geq &\frac{1}{\sqrt{1 + g_T^2} + g_T} \exp\left(\frac{1}{2}\frac{\kappa_T - 1}{\kappa_T}(g_T + \frac{1}{g_T})^2 - \frac{1}{2} g_T^2 \right) \sqrt{\frac{\kappa_T - 1}{\kappa_T}} \\
&\times \int_{g_T}^{+\infty} \frac{1}{\sqrt{2\pi}\sqrt{\frac{\kappa_T - 1}{\kappa_T}}} \exp\left(- \frac{\Big(s + (g_T + \frac{1}{g_T})\frac{\kappa_T - 1}{\kappa_T}\Big)^2}{2\frac{\kappa_T - 1}{\kappa_T}} \right) ds \\
\geq &\frac{1}{\sqrt{1 + g_T^2} + g_T} \exp\left(\frac{1}{2}\frac{\kappa_T - 1}{\kappa_T}(g_T + \frac{1}{g_T})^2 - \frac{1}{2} g_T^2 \right) \sqrt{\frac{\kappa_T - 1}{\kappa_T}} \\
&\times \left(1 - \Phi\Big(g_T\sqrt{\frac{\kappa_T}{\kappa_T - 1}} + (g_T + \frac{1}{g_T})\sqrt{\frac{\kappa_T - 1}{\kappa_T}}\Big) \right).
\end{aligned}$$

Also by Komatu's lower bound,

$$\begin{aligned}
1 - \Phi\Big(g_T\sqrt{\frac{\kappa_T}{\kappa_T - 1}} + (g_T + \frac{1}{g_T})\sqrt{\frac{\kappa_T - 1}{\kappa_T}}\Big) \geq \frac{1}{\sqrt{2\pi}} \exp\left(- \frac{1}{2}\Big(g_T\sqrt{\frac{\kappa_T}{\kappa_T - 1}} + (g_T + \frac{1}{g_T})\sqrt{\frac{\kappa_T - 1}{\kappa_T}}\Big)^2 \right) \\
\frac{2}{\sqrt{4 + \Big(g_T\sqrt{\frac{\kappa_T}{\kappa_T - 1}} + (g_T + \frac{1}{g_T})\sqrt{\frac{\kappa_T - 1}{\kappa_T}}\Big)^2} + \Big(g_T\sqrt{\frac{\kappa_T}{\kappa_T - 1}} + (g_T + \frac{1}{g_T})\sqrt{\frac{\kappa_T - 1}{\kappa_T}}\Big)}.
\end{aligned}$$

Define

$$\begin{aligned}\mathcal{C}(T) =& \Big[\sqrt{4+\Big(g_T\sqrt{\frac{\kappa_T}{\kappa_T-1}}+(g_T+\frac{1}{g_T})\sqrt{\frac{\kappa_T-1}{\kappa_T}}\Big)^2}+g_T\sqrt{\frac{\kappa_T}{\kappa_T-1}}+(g_T+\frac{1}{g_T})\sqrt{\frac{\kappa_T-1}{\kappa_T}}\Big] \\ & \times \pi\sqrt{\kappa_T}\Big(\sqrt{1+g_T^2}+g_T\Big).\end{aligned}$$

Then we get

$$\begin{aligned}\Pr(\mathcal{E}_{k^*}(x_1,x_2)) \geq& \, \mathcal{C}(T)^{-1}\left[1-\exp\Big(-2(1-\frac{1}{\sqrt{\kappa_T-1}})\frac{\sqrt{2\ln 1/\delta}}{\kappa_T-2}g_T\Big)\right] \\ & \exp\left(-\frac{1}{2}\big(g_T\sqrt{\frac{\kappa_T}{\kappa_T-1}}+(g_T+\frac{1}{g_T})\sqrt{\frac{\kappa_T-1}{\kappa_T}}\big)^2+\frac{1}{2}\frac{\kappa_T-1}{\kappa_T}\big(g_T+\frac{1}{g_T}\big)^2-\frac{1}{2}g_T^2\right) \\ \geq& \, \mathcal{C}(T)^{-1}\left[1-\exp\Big(-2(1-\frac{1}{\sqrt{\kappa_T-1}})\frac{\sqrt{2\ln 1/\delta}}{\kappa_T-2}g_T\Big)\right]\exp\Big(-1-(2+\frac{1}{2(\kappa_T-1)})g_T^2\Big).\end{aligned}$$

Now we turn to lower bounding $\Pr(\mathcal{E}_\beta(x_1,x_2))$.

$$\begin{aligned}&\Pr(\mathcal{E}_\beta(x_1,x_2)) \\ =& \Pr\Big(x_2 < \text{UCB}_\beta(a_\beta) < x_1, \min_{n\in[b_\beta]}\text{UCB}_\beta(n) > x_2\Big) \\ =& \int_{a_\beta(x_2-\mu_\beta)-\sqrt{2a_\beta\ln 1/\delta}}^{a_\beta(x_1-\mu_\beta)-\sqrt{2a_\beta\ln 1/\delta}} \Pr\Big(\min_{n\in[a_\beta]}\text{UCB}_\beta(n) > x_2, \min_{n=a_\beta+1,\ldots,b_\beta}\text{UCB}_\beta(n) > x_2\Big|S_{a_\beta}^{(\beta)}=s\Big)f(S_{a_\beta}^{(\beta)}=s)ds.\end{aligned}$$

Since $\min_{n\in[a_\beta]}\text{UCB}_\beta(n) > x_2$ and $\min_{n=a_\beta+1,\ldots,b_\beta}\text{UCB}_\beta(n) > x_2$ are independent conditioned on $S_{a_\beta}^{(\beta)}$, we have

$$\begin{aligned}&\Pr\Big(\min_{n\in[a_\beta]}\text{UCB}_\beta(n) > x_2, \min_{n=a_\beta+1,\ldots,b_\beta}\text{UCB}_\beta(n) > x_2\Big|S_{a_\beta}^{(\beta)}=s\Big) \\ =& \Pr\Big(\min_{n\in[a_\beta]}\text{UCB}_\beta(n) > x_2\Big|S_{a_\beta}^{(\beta)}=s\Big)\times \\ & \Pr\Big(\min_{n=a_\beta+1,\ldots,b_\beta}\text{UCB}_\beta(n) > x_2\Big|S_{a_\beta}^{(\beta)}=s\Big).\end{aligned}$$

We first derive a lower bound for $\Pr\Big(\min_{n\in[a_\beta]}\text{UCB}_\beta(n) > x_2\Big|S_{a_\beta}^{(\beta)}=s\Big)$ as we did for $\Pr\Big(\min_{n\in[\kappa_T-1]}\text{UCB}_{k^*}(n) > x_1\Big|S_{\kappa_T-1}^{(k^*)}=s\Big)$.

$$\begin{aligned}&\Pr\Big(\min_{n\in[a_\beta]}\text{UCB}_\beta(n) > x_2\Big|S_{a_\beta}^{(\beta)}=s\Big) \\ =& \Pr\Big(\forall n=1,\ldots,a_\beta: S_n^{(\beta)} > n(x_2-\mu_\beta)-\sqrt{2n\ln 1/\delta}|S_{a_\beta}^{(\beta)}=s\Big) \\ \geq& \Pr\Big(\forall n=1,\ldots,a_\beta: S_n^{(\beta)} > [(x_2-\mu_\beta)-\frac{1}{\sqrt{a_\beta}+1}\sqrt{2\ln 1/\delta}]n-(1-\frac{1}{\sqrt{a_\beta}+1})\sqrt{2\ln 1/\delta}\Big|S_{a_\beta}^{(\beta)}=s\Big) \\ \geq& 1-\exp\Big(-2(1-\frac{1}{\sqrt{a_\beta}+1})\sqrt{2\ln 1/\delta}\big[-a_\beta(x_2-\mu_\beta)+\sqrt{2a_\beta\ln 1/\delta}+s\big]a_\beta^{-1}\Big),\end{aligned}$$

where the last inequality is by Lemma D.3. Then we consider $\Pr\Big(\min_{n=a_\beta+1,...,b_\beta} \text{UCB}_\beta(n) > x_2 \Big| S_{a_\beta}^{(\beta)} = s\Big)$,

$$
\begin{aligned}
&\Pr\Big(\min_{n=a_\beta+1,...,b_\beta} \text{UCB}_\beta(n) > x_2 \Big| S_{a_\beta}^{(\beta)} = s\Big) \\
=&\Pr\Big(\forall n = a_\beta+1,...,b_\beta : S_{a_\beta}^{(\beta)} + \sum_{j=1+a_\beta}^{n} (X_{\beta,j} - \mu_\beta) > n(x_2 - \mu_\beta) - \sqrt{2n\ln 1/\delta} \Big| S_{a_\beta}^{(\beta)} = s\Big) \\
=&\Pr\Big(\forall n = a_\beta+1,...,b_\beta : s + \sum_{j=1+a_\beta}^{n} (X_{\beta,j} - \mu_\beta) > n(x_2 - \mu_\beta) - \sqrt{2n\ln 1/\delta}\Big) \\
=&\Pr\Big(\forall n = 1,...,b_\beta - a_\beta : s + S_n^{(\beta)} > (n + a_\beta)(x_2 - \mu_\beta) - \sqrt{2(n + a_\beta)\ln 1/\delta}\Big) \\
\geq&\Pr\Big(\forall n = 1,...,b_\beta - a_\beta : S_n^{(\beta)} > \frac{(b_\beta - a_\beta)(x_2 - \mu_\beta) - \sqrt{2\ln 1/\delta}(\sqrt{b_\beta} - \sqrt{a_\beta})}{b_\beta - a_\beta} n \\
&+ a_\beta(x_2 - \mu_\beta) - \sqrt{2a_\beta \ln 1/\delta - s}\Big) \\
=&\Pr\Big(\forall n = 1,...,b_\beta - a_\beta : S_n^{(\beta)} > \big(x_2 - \mu_\beta - \frac{\sqrt{2\ln 1/\delta}}{\sqrt{b_\beta} + \sqrt{a_\beta}}\big)n + a_\beta(x_2 - \mu_\beta) - \sqrt{2a_\beta \ln 1/\delta - s}\Big) \\
=&\Big[1 - \Pr\Big(\exists n = 1,...,b_\beta - a_\beta : S_n^{(\beta)} \leq \big(x_2 - \mu_\beta - \frac{\sqrt{2\ln 1/\delta}}{\sqrt{b_\beta} + \sqrt{a_\beta}}\big)n + a_\beta(x_2 - \mu_\beta) - \sqrt{2a_\beta \ln 1/\delta - s}\Big)\Big],
\end{aligned}
$$

where the inequality is due to convexity. We also note that, the joint distribution of $S_1^{(\beta)}, ..., S_{b_\beta - a_\beta}^{(\beta)}$ is identical to that of $B(1), ..., B(b_\beta - a_\beta)$. Thus we can further lower bound $\Pr(\mathcal{E}_\beta(x_1, x_2))$ as

$$
\begin{aligned}
&\Pr(\mathcal{E}_\beta(x_1, x_2)) \\
\geq&\int_{a_\beta(x_2-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}}^{a_\beta(x_1-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}} \Pr\Big(\min_{n\in[a_\beta]} \text{UCB}_\beta(n) > x_2 \Big| S_{a_\beta}^{(\beta)} = s\Big)\Big[1 - \Pr\Big(\exists t \in [0, b_\beta - a_\beta] : \\
&B(t) \leq \big(x_2 - \mu_\beta - \frac{\sqrt{2\ln 1/\delta}}{\sqrt{b_\beta} + \sqrt{a_\beta}}\big)t + a_\beta(x_2 - \mu_\beta) - \sqrt{2a_\beta \ln 1/\delta - s}\Big)\Big] f(S_{a_\beta}^{(\beta)} = s)ds \\
\geq&\Big[1 - \exp\Big(-(1 - \frac{1}{\sqrt{a_\beta} + 1})\sqrt{2\ln 1/\delta}\frac{g_T}{\kappa_T}\Big)\Big] \int_{a_\beta(\frac{x_1+x_2}{2}-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}}^{a_\beta(x_1-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}} \Big[1- \\
&\Pr\Big(\exists t \in [0, b_\beta - a_\beta] : B(t) \leq \big(x_2 - \mu_\beta - \frac{\sqrt{2\ln 1/\delta}}{\sqrt{b_\beta} + \sqrt{a_\beta}}\big)t + a_\beta(x_2 - \mu_\beta) - \sqrt{2a_\beta \ln 1/\delta - s}\Big)\Big] f(S_{a_\beta}^{(\beta)} = s)ds \\
=:&\Big[1 - \exp\Big(-(1 - \frac{1}{\sqrt{a_\beta} + 1})\sqrt{2\ln 1/\delta}\frac{g_T}{\kappa_T}\Big)\Big] P_\beta.
\end{aligned}
$$

Now it suffices to lower bound P_β. Again, here we need a proposition in Scheike (1992) that characterizes the probability of Brownian motion crossing a linear boundary within a finite time horizon.

Lemma D.4 (Proposition 2, Scheike (1992))**.** *If $B(t)$ is standard Brownian motion, and $a > 0$, $T < \infty$, then*

$$
\Pr\Big(\bigcup_{t\in[0,T]} (B(t) \geq a + bt)\Big) = 1 - \Phi(\frac{a}{\sqrt{T}} + b\sqrt{T}) + \exp(-2ab)\Phi(-\frac{a}{\sqrt{T}} + b\sqrt{T}).
$$

If $a \leq 0$, the probability is 1.

By setting

$$
a = s - a_\beta(x_2 - \mu_\beta) + \sqrt{2a_\beta \ln 1/\delta}, \quad b = -\big(x_2 - \mu_\beta - \frac{\sqrt{2\ln 1/\delta}}{\sqrt{b_\beta} + \sqrt{a_\beta}}\big),
$$

we have that

$$
\begin{aligned}
P_\beta \geq \int_{a_\beta(\frac{x_1+x_2}{2}-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}}^{a_\beta(x_1-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}} & \Big[\Phi\Big(\frac{s-a_\beta(x_2-\mu_\beta)+\sqrt{2a_\beta \ln 1/\delta}}{\sqrt{b_\beta - a_\beta}} - \big(x_2-\mu_\beta-\frac{\sqrt{2\ln 1/\delta}}{\sqrt{b_\beta}+\sqrt{a_\beta}}\big)\sqrt{b_\beta-a_\beta}\Big) \\
& - \exp\Big(2\big(x_2-\mu_\beta-\frac{\sqrt{2\ln 1/\delta}}{\sqrt{b_\beta}+\sqrt{a_\beta}}\big)\big(s-a_\beta(x_2-\mu_\beta)+\sqrt{2a_\beta \ln 1/\delta}\big)\Big) \\
& \times \Phi\Big(-\frac{s-a_\beta(x_2-\mu_\beta)+\sqrt{2a_\beta \ln 1/\delta}}{\sqrt{b_\beta-a_\beta}} - \big(x_2-\mu_\beta-\frac{\sqrt{2\ln 1/\delta}}{\sqrt{b_\beta}+\sqrt{a_\beta}}\big)\sqrt{b_\beta-a_\beta}\Big)\Big] f(S_{a_\beta}^{(\beta)}=s)ds.
\end{aligned}
$$

For notational convenience, we denote

$$
A = \big(x_2-\mu_\beta-\frac{\sqrt{2\ln 1/\delta}}{\sqrt{b_\beta}+\sqrt{a_\beta}}\big)\sqrt{b_\beta-a_\beta}, \quad B = \frac{s-a_\beta(x_2-\mu_\beta)+\sqrt{2a_\beta \ln 1/\delta}}{\sqrt{b_\beta-a_\beta}}.
$$

For $\lambda, \eta, a_1, a_2 \in (0,1)$, we set

$$
a_\beta = \left\lfloor (1-\eta)\frac{2\ln 1/\delta}{(x_1-\mu_\beta)^2} \right\rfloor, \quad b_\beta = \left\lceil (1+\lambda)\frac{2\ln 1/\delta}{(x_2-\mu_\beta)^2} \right\rceil,
$$

$$
g_T = (\ln T)^{a_1}, \quad \kappa_T = 3 + \lceil (\ln T)^{a_2} \rceil.
$$

We require that a_1, a_2 satisfy the following conditions:

1. $a_2 - \frac{1}{2} < \frac{a_2}{2} < a_1 < \frac{1}{2}$
2. $a_1 + \frac{a_2}{2} > \frac{1}{2}$

In the below, we take $T \to +\infty$ and evaluate the asymptotic rates of a_β, b_β, A, B.

$$
\lim_{T\to+\infty} \frac{a_\beta}{(\ln T)^{a_2}} = (1-\eta) \lim_{T\to+\infty} \frac{1}{(\ln T)^{a_2}} \frac{4\ln T}{\Big(\Delta_\beta + \sqrt{\frac{4\ln T}{\kappa_T - 1}}\Big)^2} = (1-\eta) \lim_{T\to+\infty} \frac{\kappa_T - 1}{(\ln T)^{a_2}} = 1-\eta,
$$

$$
\lim_{T\to+\infty} \frac{b_\beta}{(\ln T)^{a_2}} = (1+\lambda) \lim_{T\to+\infty} \frac{1}{(\ln T)^{a_2}} \frac{4\ln T}{\Big(\Delta_\beta + \sqrt{\frac{4\ln T}{\kappa_T}} - \frac{g_T}{\kappa_T}\Big)^2}.
$$

Since $\frac{1}{2} + \frac{1}{2}a_2 > \frac{1}{2} > a_1$ by condition 1., we further have

$$
\lim_{T\to+\infty} \frac{b_\beta}{(\ln T)^{a_2}} = (1+\lambda) \lim_{T\to+\infty} \frac{1}{(\ln T)^{a_2}} \frac{4\ln T}{\Big(\Delta_\beta + \sqrt{\frac{4\ln T}{\kappa_T}}\Big)^2} = 1+\lambda,
$$

$$
\lim_{T\to+\infty} \frac{b_\beta - a_\beta}{(\ln T)^{a_2}} = (1+\lambda) - (1-\eta) = \lambda + \eta,
$$

$$
\begin{aligned}
\lim_{T\to+\infty} \frac{A}{(\ln T)^{\frac{1}{2}}} &= \lim_{T\to+\infty} \frac{1}{(\ln T)^{\frac{1}{2}}} \big(\Delta_\beta + \sqrt{\frac{4\ln T}{\kappa_T}} - \frac{g_T}{\kappa_T} - \frac{\sqrt{4\ln T}}{\sqrt{b_\beta}+\sqrt{a_\beta}}\big)\sqrt{b_\beta - a_\beta} \\
&= \lim_{T\to+\infty} \frac{1}{(\ln T)^{\frac{1}{2}}} \big(\sqrt{4\ln T}(\frac{1}{\sqrt{\kappa_T}} - \frac{1}{\sqrt{b_\beta}+\sqrt{a_\beta}}) - \frac{g_T}{\kappa_T}\big)\sqrt{b_\beta - a_\beta},
\end{aligned}
$$

we observe that $\lim_{T\to+\infty}\Big(\frac{1}{\sqrt{\kappa_T}} - \frac{1}{\sqrt{b_\beta}+\sqrt{a_\beta}}\Big)\sqrt{4\ln T}/(\ln T)^{\frac{1}{2}-\frac{a_2}{2}} = 2\Big(1 - \frac{1}{\sqrt{1+\lambda}+\sqrt{1-\eta}}\Big) > 0$, $\lim_{T\to+\infty}\frac{g_T}{\kappa_T}/(\ln T)^{a_1-a_2} = 1 > 0$, and again $\frac{1}{2}-\frac{a_2}{2} > a_1 - a_2$, thus

$$\begin{aligned}\lim_{T\to+\infty}\frac{A}{(\ln T)^{\frac{1}{2}}} &= \lim_{T\to+\infty}\frac{1}{(\ln T)^{\frac{1}{2}}}\big(\sqrt{4\ln T}(\frac{1}{\sqrt{\kappa_T}} - \frac{1}{\sqrt{b_\beta}+\sqrt{a_\beta}})\big)\sqrt{b_\beta - a_\beta}\\ &= 2\sqrt{\lambda+\eta}\Big(1-\frac{1}{\sqrt{1+\lambda}+\sqrt{1-\eta}}\Big).\end{aligned}$$

Finally we consider B. Note that B depends on the value of s, and in the integral $s \in \Big[a_\beta(\frac{x_1+x_2}{2}-\mu_\beta) - \sqrt{2a_\beta\ln 1/\delta}, a_\beta(x_1-\mu_\beta) - \sqrt{2a_\beta \ln 1/\delta}\Big]$. Thus B is bounded by

$$0 < \frac{a_\beta(x_1-x_2)}{2\sqrt{b_\beta - a_\beta}} \le B \le \frac{a_\beta(x_1-x_2)}{\sqrt{b_\beta-a_\beta}}.$$

Now it suffices to evaluate

$$\lim_{T\to+\infty}\frac{1}{(\ln T)^{a_1-\frac{a_2}{2}}}\frac{a_\beta(x_1-x_2)}{\sqrt{b_\beta-a_\beta}} = \lim_{T\to+\infty}\frac{1}{(\ln T)^{a_1-\frac{a_2}{2}}}\frac{a_\beta}{\sqrt{b_\beta-a_\beta}}\Big(\frac{g_T}{\kappa_T}+\sqrt{\frac{4\ln T}{\kappa_T-1}}-\sqrt{\frac{4\ln T}{\kappa_T}}\Big).$$

We observe that $\lim_{T\to+\infty}\Big(\sqrt{\frac{4\ln T}{\kappa_T-1}}-\sqrt{\frac{4\ln T}{\kappa_T}}\Big)/(\ln T)^{\frac{1}{2}-\frac{3}{2}a_2} = 1 > 0$ and that by condition 2, for a_1, a_2,

$$\lim_{T\to+\infty}\frac{1}{(\ln T)^{a_1-\frac{a_2}{2}}}\frac{a_\beta(x_1-x_2)}{\sqrt{b_\beta-a_\beta}} = \lim_{T\to+\infty}\frac{1}{(\ln T)^{a_1-\frac{a_2}{2}}}\frac{a_\beta}{\sqrt{b_\beta-a_\beta}}\frac{g_T}{\kappa_T} = \frac{1-\eta}{\sqrt{\lambda+\eta}}.$$

Comparing A and B, since $\frac{1}{2} > a_1 - \frac{a_2}{2}$, we have that $\exists T_0 \in \mathbb{N}^+$: $\forall T > T_0$, $A > B$. And we also have that $\forall s \in \Big[a_\beta(\frac{x_1+x_2}{2}-\mu_\beta) - \sqrt{2a_\beta\ln 1/\delta}, a_\beta(x_1-\mu_\beta) - \sqrt{2a_\beta\ln 1/\delta}\Big]$

$$\begin{aligned}&\frac{2}{\sqrt{4+(A-B)^2}+(A-B)} - \frac{2}{\sqrt{2+(A+B)^2}+(A+B)}\\ \ge&\frac{2}{\sqrt{4+\Big(A-\frac{a_\beta(x_1-x_2)}{2\sqrt{b_\beta-a_\beta}}\Big)^2}+\Big(A-\frac{a_\beta(x_1-x_2)}{2\sqrt{b_\beta-a_\beta}}\Big)} - \frac{2}{\sqrt{2+\Big(A+\frac{a_\beta(x_1-x_2)}{2\sqrt{b_\beta-a_\beta}}\Big)^2}+\Big(A+\frac{a_\beta(x_1-x_2)}{2\sqrt{b_\beta-a_\beta}}\Big)}.\end{aligned}$$

Denoting $\underline{B} = \frac{a_\beta(x_1-x_2)}{2\sqrt{b_\beta-a_\beta}}$, we have that $\exists T_1 \in \mathbb{N}$, $\forall T > T_1$, $A\underline{B} > 1/2$. By straightforward calculation, we can verify that for any $T > T_1$,

$$\frac{2}{\sqrt{4+(A-\underline{B})^2}+(A-\underline{B})} - \frac{2}{\sqrt{2+(A+\underline{B})^2}+(A+\underline{B})} > 0.$$

Moreover,

$$\begin{aligned}&\lim_{T\to+\infty}\left[\frac{2}{\sqrt{4+(A-\underline{B})^2}+(A-\underline{B})} - \frac{2}{\sqrt{2+(A+\underline{B})^2}+(A+\underline{B})}\right]\frac{A^2}{\underline{B}}\\ =&\lim_{T\to+\infty}\frac{\sqrt{2+(A+\underline{B})^2}+A+\underline{B}-\sqrt{4+(A-\underline{B})^2}-A+\underline{B}}{2\underline{B}}\\ =&1+\lim_{T\to+\infty}\frac{\sqrt{2+(A+\underline{B})^2}-\sqrt{4+(A-\underline{B})^2}}{2\underline{B}} = 1+\lim_{T\to+\infty}\frac{4A\underline{B}-2}{2\underline{B}(\sqrt{2+(A+\underline{B})^2}+\sqrt{4+(A-\underline{B})^2})}\\ =&1+\lim_{T\to+\infty}\frac{4A}{2(\sqrt{2+(A+\underline{B})^2}+\sqrt{4+(A-\underline{B})^2})} = 2.\end{aligned}$$

Adopting both Komatu's lower bound and upper bound simultaneously, we again focus on P_β

$$\begin{aligned}
P_\beta \geq &\int_{a_\beta(\frac{x_1+x_2}{2}-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}}^{a_\beta(x_1-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}} \Big[\Phi(B-A) - \exp(2AB)\times\Phi(-B-A)\Big] f(S^{(\beta)}_{a_\beta}=s)ds \\
\geq &\int_{a_\beta(\frac{x_1+x_2}{2}-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}}^{a_\beta(x_1-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}} \frac{1}{\sqrt{2\pi}}\Big[\frac{2}{\sqrt{4+(A-B)^2}+(A-B)}\exp(-\frac{1}{2}(A-B)^2) \\
&- \exp(2AB)\frac{2}{\sqrt{2+(A+B)^2}+(A+B)}\exp(-\frac{1}{2}(A+B)^2)\Big] f(S^{(\beta)}_{a_\beta}=s)ds \\
= &\int_{a_\beta(\frac{x_1+x_2}{2}-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}}^{a_\beta(x_1-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}} \frac{1}{\sqrt{2\pi}}\Big[\frac{2}{\sqrt{4+(A-B)^2}+(A-B)} \\
&- \frac{2}{\sqrt{2+(A+B)^2}+(A+B)}\Big]\exp(-\frac{1}{2}(A-B)^2) f(S^{(\beta)}_{a_\beta}=s)ds.
\end{aligned}$$

When $T > \max\{T_0, T_1\}$, noting that $S^{(\beta)}_{a_\beta} \sim \mathcal{N}(0, a_\beta)$,

$$\begin{aligned}
P_\beta \geq &\Big[\frac{2}{\sqrt{4+(A-\underline{B})^2}+(A-\underline{B})} - \frac{2}{\sqrt{2+(A+\underline{B})^2}+(A+\underline{B})}\Big] \\
&\int_{a_\beta(\frac{x_1+x_2}{2}-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}}^{a_\beta(x_1-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}} \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(A-B)^2) f(S^{(\beta)}_{a_\beta}=s)ds \\
\geq &\Big[\frac{2}{\sqrt{4+(A-\underline{B})^2}+(A-\underline{B})} - \frac{2}{\sqrt{2+(A+\underline{B})^2}+(A+\underline{B})}\Big] \\
&\int_{a_\beta(\frac{x_1+x_2}{2}-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}}^{a_\beta(x_1-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta}} \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}A^2)\frac{1}{\sqrt{2\pi a_\beta}}\exp(-\frac{s^2}{2a_\beta})ds.
\end{aligned}$$

Since the upper limit of the integral is negative, i.e.

$$a_\beta(x_1-\mu_\beta) - \sqrt{2a_\beta \ln 1/\delta} \leq (1-\eta-\sqrt{1-\eta})\frac{2\ln 1/\delta}{x_1-\mu_\beta} < 0, \quad \eta \in (0,1),$$

we have that

$$\begin{aligned}
&\exp(-\frac{1}{2}A^2 - \frac{s^2}{2a_\beta}) \\
\geq &\exp\Big(-\frac{1}{2}A^2 - \frac{(a_\beta(\frac{x_1+x_2}{2}-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta})^2}{2a_\beta}\Big) \\
> &\exp\Big(-\frac{1}{2}A^2 - \frac{(a_\beta(x_2-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta})^2}{2a_\beta}\Big) \\
\geq &\exp\Big(-\frac{1}{2}\big[x_2-\mu_\beta-\frac{\sqrt{2\ln 1/\delta}}{\sqrt{b_\beta}+\sqrt{a_\beta}}\big]^2(b_\beta-a_\beta) - \frac{1}{2}\big[x_2-\mu_\beta-\frac{\sqrt{2\ln 1/\delta}}{\sqrt{a_\beta}}\big]^2 a_\beta\Big).
\end{aligned}$$

Now we can get rid of the integral and rewrite the lower bound for P_β as

$$
\begin{aligned}
P_\beta \geq & \Big[\frac{2}{\sqrt{4+(A-\underline{B})^2}+(A-\underline{B})} - \frac{2}{\sqrt{2+(A+\underline{B})^2}+(A+\underline{B})} \Big] \\
& \exp\Big(-\frac{1}{2}\big[x_2-\mu_\beta-\frac{\sqrt{2\ln 1/\delta}}{\sqrt{b_\beta}+\sqrt{a_\beta}}\big]^2(b_\beta-a_\beta) - \frac{1}{2}\big[x_2-\mu_\beta-\frac{\sqrt{2\ln 1/\delta}}{\sqrt{a_\beta}}\big]^2 a_\beta \Big) \\
& \frac{1}{2\pi\sqrt{a_\beta}}\Big(a_\beta(x_1-\mu_\beta)-\sqrt{2a_\beta \ln 1/\delta} - a_\beta(\frac{x_1+x_2}{2}-\mu_\beta)+\sqrt{2a_\beta\ln 1/\delta} \Big) \\
= & \Big[\frac{2}{\sqrt{4+(A-\underline{B})^2}+(A-\underline{B})} - \frac{2}{\sqrt{2+(A+\underline{B})^2}+(A+\underline{B})} \Big] \\
& \exp\Big(-\frac{1}{2}\big[x_2-\mu_\beta-\frac{\sqrt{2\ln 1/\delta}}{\sqrt{b_\beta}+\sqrt{a_\beta}}\big]^2(b_\beta-a_\beta) - \frac{1}{2}\big[x_2-\mu_\beta-\frac{\sqrt{2\ln 1/\delta}}{\sqrt{a_\beta}}\big]^2 a_\beta \Big) \frac{(x_1-x_2)\sqrt{a_\beta}}{4\pi} =: \text{LB},
\end{aligned}
$$

if $T > \max\{T_0, T_1\}$. Since we have shown that events $\mathcal{E}_{k^*}$ and $\mathcal{E}_\beta$ happen with non-negligible probability, now we focus on the expected regret lower bound of UCB algorithm,

$$
\begin{aligned}
R_T = & \mathbb{E}\big[\sum_{t=1}^{T}\Delta_{a_t}\big] = \mathbb{E}\big[\sum_{t=1}^{T}\Delta_{a_t}(\mathbb{I}(\mathcal{E}_{k^*}\cap\mathcal{E}_\beta)+\mathbb{I}(\bar{\mathcal{E}}_{k^*}\cup\bar{\mathcal{E}}_\beta))\big] \\
\geq & \mathbb{E}\big[\sum_{t=1}^{T}\Delta_{a_t}\mathbb{I}(\mathcal{E}_{k^*}\cap\mathcal{E}_\beta)\big] = \mathbb{E}\big[\mathbb{I}(\mathcal{E}_{k^*}\cap\mathcal{E}_\beta)\sum_{i\neq k^*}^{K}\Delta_i\sum_{t=1}^{T}\mathbb{I}(a_t=i)\big] \\
\geq & \Delta_{\min}\mathbb{E}\big[\mathbb{I}(\mathcal{E}_{k^*}\cap\mathcal{E}_\beta)\sum_{i\neq k^*}^{K}\sum_{t=1}^{T}\mathbb{I}(a_t=i)\big] = \Delta_{\min}\mathbb{E}\big[\mathbb{I}(\mathcal{E}_{k^*}\cap\mathcal{E}_\beta)\sum_{t=1}^{T}\mathbb{I}(a_t\neq k^*)\big].
\end{aligned}
$$

Recall that by assumption, $m_{k^*} = O((\ln T)^\theta), \theta \in (0,1)$. Given this θ, $\exists \epsilon > 0$: $\max\{\frac{1}{2},\theta\}+2\epsilon < 1$. We set $a_1 = \frac{1}{2}\max\{\frac{1}{2},\theta\}+\epsilon$, $a_2 = \max\{\frac{1}{2},\theta\}+\epsilon$. Obviously $a_1, a_2 \in (0,1)$. Here we validate the conditions that a_1, a_2 are required to satisfy:

1. $$a_2 - \frac{1}{2} < \frac{a_2}{2} = \frac{1}{2}\max\{\frac{1}{2},\theta\}+\frac{1}{2}\epsilon < a_1 = \frac{1}{2}\max\{\frac{1}{2},\theta\}+\epsilon < \frac{1}{2}.$$

2. $$a_1 + \frac{a_2}{2} = \max\{\frac{1}{2},\theta\}+\frac{3}{2}\epsilon > \max\{\frac{1}{2},\theta\} \geq \frac{1}{2}.$$

Fix any $\gamma \in (0,1)$, we set

$$\lambda = \frac{\gamma}{16}, \quad \eta = \min\Big\{\frac{\gamma}{32}, 1-\frac{1}{(1+\sqrt{\gamma}/4\sqrt{2})^2}\Big\}.$$

Also obviously, $\lambda, \eta \in (0,1)$. We now show that, under the event $\mathcal{E}_{k^*}\cap\mathcal{E}_\beta$ and when T is large enough, the optimal arm is pulled at most κ_T times before leaving. Under both events, when the number of pulls of arm k^* is smaller than κ_T, (1) arm k^*'s UCB index is greater than x_1 and (2) the number of pulls of arm β is no greater than a_β. When the number of pulls of arm k^* is $\kappa_T - 1$ and at some time arm k^*'s UCB index is the largest among all active arms, it will be pulled the κ_T-th time. After that, arm k^*'s UCB index is below x_2. By $\mathcal{E}_\beta$, we know that before arm k^*'s UCB index becomes the largest again, at least the $(a_\beta+1)$-th, $(a_\beta+2)$-th, ..., $(b_\beta+1)$-th pulls of arm β need to be done. The total number of sub-optimal pulls (contributed by arm β) after arm k^* is pulled the κ_T-th time is at least $b_\beta - a_\beta + 1$, not to mention the arms other than k^* and β. By our assignment, $a_2 > \theta$. Besides, we have shown that $\lim_{T\to+\infty}\frac{b_\beta-a_\beta}{(\ln T)^{a_2}} = \lambda+\eta > 0$. As a result,

$$\lim_{T\to+\infty}\frac{m_{k^*}}{b_\beta-a_\beta+1} = \lim_{T\to+\infty}\frac{m_{k^*}}{(\ln T)^{a_2}}\frac{(\ln T)^{a_2}}{b_\beta-a_\beta+1} = 0.$$

Thus $\exists T_2 \in \mathbb{N}^+, \forall T > T_2, m_{k^*} < b_\beta - a_\beta + 1$. Under the event $\mathcal{E}_{k^*} \cap \mathcal{E}_\beta$, when $T > T_2$, arm k^* is pulled at most κ_T times. We obtain that when $T > T_2$,

$$\begin{aligned} R_T \geq & \Delta_{\min} \mathbb{E}\big[\mathbb{I}(\mathcal{E}_{k^*} \cap \mathcal{E}_\beta) \sum_{t=1}^{T} \mathbb{I}(a_t \neq k^*)\big] \geq \Delta_{\min} \mathbb{E}\big[\mathbb{I}(\mathcal{E}_{k^*} \cap \mathcal{E}_\beta)\big](T - \kappa_T) \\ & = \Delta_{\min} \Pr(\mathcal{E}_{k^*} \cap \mathcal{E}_\beta)(T - \kappa_T) = \Delta_{\min} \Pr(\mathcal{E}_{k^*}) \Pr(\mathcal{E}_\beta)(T - \kappa_T), \end{aligned}$$

noticing that $\mathcal{E}_{k^*} \perp \mathcal{E}_\beta \mid x_1, x_2$. For $\mathcal{E}_{k^*}$, we have

$$0 \leq \lim_{T \to +\infty} \frac{T^{-\frac{1}{2}\gamma} \mathcal{C}(T)^{-1}}{\Pr(\mathcal{E}_{k^*})} \leq \lim_{T \to +\infty} \frac{T^{-\frac{1}{2}\gamma} \mathcal{C}(T)^{-1}}{\mathcal{C}(T)^{-1}} \exp\Big(1 + (2 + \frac{1}{2(\kappa_T - 1)}) g_T^2\Big),$$

where we have used the fact that

$$\lim_{T \to +\infty} 1 - \exp\Big(-2(1 - \frac{1}{\sqrt{\kappa_T - 1}}) \frac{\sqrt{2 \ln 1/\delta}}{\kappa_T - 2} g_T\Big) = 1,$$

since $\frac{1}{2} + a_1 - a_2 > 0$. Furthermore,

$$0 \leq \lim_{T \to +\infty} \frac{T^{-\frac{1}{2}\gamma} \mathcal{C}(T)^{-1}}{\Pr(\mathcal{E}_{k^*})} \leq \lim_{T \to +\infty} \frac{\mathcal{C}(T)^{-1}}{\mathcal{C}(T)^{-1}} \exp\Big(1 + (2 + \frac{1}{2(\kappa_T - 1)}) g_T^2 - \frac{\gamma}{2} \ln T\Big) = 0,$$

because $2a_1 < 1$. Thus $\lim_{T \to +\infty} \frac{T^{-\frac{1}{2}\gamma} \mathcal{C}(T)^{-1}}{\Pr(\mathcal{E}_{k^*})} = 0$, $\Pr(\mathcal{E}_{k^*}) = \Omega(T^{-\frac{1}{2}\gamma} \mathcal{C}(T)^{-1})$. For $\mathcal{E}_\beta$, since $P_\beta \geq \text{LB}, \forall T > \max\{T_0, T_1\}$, define

$$h(T) = \frac{1}{2}\big[x_2 - \mu_\beta - \frac{\sqrt{2 \ln 1/\delta}}{\sqrt{b_\beta} + \sqrt{a_\beta}}\big]^2 (b_\beta - a_\beta) + \frac{1}{2}\big[x_2 - \mu_\beta - \frac{\sqrt{2 \ln 1/\delta}}{\sqrt{a_\beta}}\big]^2 a_\beta,$$

we note that

$$\begin{aligned} & \lim_{T \to +\infty} \frac{h(T)}{\ln T} \\ = & \lim_{T \to +\infty} \frac{1}{2} \frac{b_\beta - a_\beta}{(\ln T)^{a_2}} \Big[\Delta_\beta (\ln T)^{\frac{a_2}{2} - \frac{1}{2}} + 2(\ln T)^{\frac{a_2}{2} - \frac{1}{2}} \sqrt{\frac{\ln T}{\kappa_T}} - \frac{g_T}{\kappa_T} (\ln T)^{\frac{a_2}{2} - \frac{1}{2}} - \frac{2}{\sqrt{\frac{b_\beta}{(\ln T)^{a_2}}} + \sqrt{\frac{a_\beta}{(\ln T)^{a_2}}}}\Big]^2 \\ & + \frac{1}{2} \frac{a_\beta}{(\ln T)^{a_2}} \Big[\Delta_\beta (\ln T)^{\frac{a_2}{2} - \frac{1}{2}} + 2(\ln T)^{\frac{a_2}{2} - \frac{1}{2}} \sqrt{\frac{\ln T}{\kappa_T}} - \frac{g_T}{\kappa_T} (\ln T)^{\frac{a_2}{2} - \frac{1}{2}} - \frac{2}{\sqrt{\frac{a_\beta}{(\ln T)^{a_2}}}}\Big]^2 \\ = & \lim_{T \to +\infty} \frac{1}{2} \frac{b_\beta - a_\beta}{(\ln T)^{a_2}} \Big[2(\ln T)^{\frac{a_2}{2} - \frac{1}{2}} \sqrt{\frac{\ln T}{\kappa_T}} - \frac{2}{\sqrt{\frac{b_\beta}{(\ln T)^{a_2}}} + \sqrt{\frac{a_\beta}{(\ln T)^{a_2}}}}\Big]^2 \\ & + \frac{1}{2} \frac{a_\beta}{(\ln T)^{a_2}} \Big[2(\ln T)^{\frac{a_2}{2} - \frac{1}{2}} \sqrt{\frac{\ln T}{\kappa_T}} - \frac{2}{\sqrt{\frac{a_\beta}{(\ln T)^{a_2}}}}\Big]^2, \end{aligned}$$

since $a_2 < 1$ and $a_1 < \frac{1}{2} < \frac{1}{2} + \frac{a_2}{2}$. Thus

$$\begin{aligned} \lim_{T \to +\infty} \frac{h(T)}{\ln T} = & \frac{1}{2}(\lambda + \eta)\Big[2 - \frac{2}{\sqrt{1 + \lambda} + \sqrt{1 - \eta}}\Big]^2 + \frac{1}{2}(1 - \eta)\Big[2 - \frac{2}{\sqrt{1 - \eta}}\Big]^2 \\ = & 2(\lambda + \eta)\Big[1 - \frac{1}{\sqrt{1 + \lambda} + \sqrt{1 - \eta}}\Big]^2 + 2(1 - \eta)\Big[1 - \frac{1}{\sqrt{1 - \eta}}\Big]^2 \\ \leq & 2(\lambda + \eta) + 2\Big[1 - \frac{1}{\sqrt{1 - \eta}}\Big]^2 \\ \leq & 2(\frac{\gamma}{16} + \frac{\gamma}{32}) + 2\frac{\gamma}{32} = \frac{\gamma}{4} < \frac{\gamma}{2}. \end{aligned}$$

Thus

$$\lim_{T\to+\infty} \exp(h(T) - \frac{\gamma}{2}\ln T) = 0.$$

For $\mathcal{E}_\beta$, we have

$$\begin{aligned}
0 &\leq \lim_{T\to+\infty} \frac{T^{-\frac{1}{2}\gamma}(\ln T)^{-1}}{\Pr(\mathcal{E}_\beta)} \leq \lim_{T\to+\infty} \frac{T^{-\frac{1}{2}\gamma}(\ln T)^{-1}}{\text{LB}} \\
&\leq \lim_{T\to+\infty} T^{-\frac{1}{2}\gamma}(\ln T)^{-1} \frac{4\pi}{(x_1-x_2)\sqrt{a_\beta}} \frac{A^2}{\frac{a_\beta(x_1-x_2)}{2\sqrt{b_\beta-a_\beta}}} \exp(h(T)) \\
&= \lim_{T\to+\infty} \frac{8\pi A^2(\ln T)^{-1}}{(x_1-x_2)^2(a_\beta)^{\frac{3}{2}}} \sqrt{b_\beta-a_\beta} \exp(h(T) - \frac{\gamma}{2}\ln T) = 0,
\end{aligned}$$

since $a_1 \geq \frac{1}{2}a_2$ and $\frac{1}{2} + a_1 - a_2 > 0$. Note that we have used

$$\lim_{T\to+\infty} 1 - \exp\Big(-(1 - \frac{1}{\sqrt{a_\beta}+1})\sqrt{2\ln 1/\delta}\frac{g_T}{\kappa_T}\Big) = 1.$$

Thus $\lim_{T\to+\infty} \frac{T^{-\frac{1}{2}\gamma}(\ln T)^{-1}}{\Pr(\mathcal{E}_\beta)} = 0$, $\Pr(\mathcal{E}_\beta) = \Omega\Big(T^{-\frac{1}{2}\gamma}(\ln T)^{-1}\Big)$. Finally, we reach a conclusion that

$$R_T = \Omega\Big(\Delta_{\min}T^{1-\gamma}\big(\mathcal{C}(T)\ln T\big)^{-1}\Big)$$

for $\forall\gamma \in (0,1)$. □

E. Proofs for SE Algorithm

Proof of Proposition 3.6. Construct a problem instance: $K = 4$, $\boldsymbol{m} = (20, 20, 3, 2)$, and $\boldsymbol{\mu} = (0, 0, 0, \bar{\Delta})$. In this case, $l(K, \boldsymbol{m}) = \frac{14}{15} < 1$. Since SE algorithm pulls $a_1 = 1, a_2 = 2$, arm 4 exits the game at the end of time 2. The optimal arm is never pulled during the whole game, thus the expected regret $R_T = \bar{\Delta}T$. □

F. Proofs for FC-SE Algorithm

Proof of Theorem 4.1. Inspired by the regret analysis of the SE algorithm in Lancewicki et al. (2021), we define τ_k as the last time the algorithm attempts to do arm elimination but arm k is NOT eliminated. Clearly, by this definition, at the end of the next feasible cycle, arm k is eliminated. Arm k is active for only one more cycle, so we have $T_k(T) \leq T_k(\tau_k) + n_k$. At the end of time τ_k, since arm k is not eliminated,

$$\hat{\mu}_{k^*,T_{k^*}(\tau_k)} - 2\sqrt{\frac{\ln T}{T_{k^*}(\tau_k)}} \leq \hat{\mu}_{k,T_k(\tau_k)} + 2\sqrt{\frac{\ln T}{T_k(\tau_k)}}. \tag{12}$$

We define the good event

$$G = \Big\{\forall t \in [T], k \in [K] : |\hat{\mu}_{k,T_k(t)} - \mu_k| \leq 2\sqrt{\frac{\ln T}{T_k(t)}}\Big\} \tag{13}$$

meaning that the empirical means are close to the true reward means for any arm at any time. Then we show that G happens with high probability by presenting an upper bound on $\Pr(\neg G)$,

$$\begin{aligned}
\Pr(\neg G) &= \Pr\left(\exists t \in [T], k \in [K] : \hat{\mu}_{k,T_k(t)} - \mu_i > 2\sqrt{\frac{\ln T}{T_k(t)}} \vee \hat{\mu}_{k,T_k(t)} - \mu_k < -2\sqrt{\frac{\ln T}{T_k(t)}}\right) \\
&\le \Pr\left(\exists n \in [T], k \in [K] : \hat{\mu}_{k,n} - \mu_k > 2\sqrt{\frac{\ln T}{n}} \vee \hat{\mu}_{k,n} - \mu_k < -2\sqrt{\frac{\ln T}{n}}\right) \\
&\le \sum_{n=1}^{T}\sum_{k=1}^{K}\left[\Pr\left(\hat{\mu}_{k,n} - \mu_k > 2\sqrt{\frac{\ln T}{n}}\right) + \Pr\left(\hat{\mu}_{k,n} - \mu_k < -2\sqrt{\frac{\ln T}{n}}\right)\right] \\
&\le \sum_{n=1}^{T}\sum_{k=1}^{K}\frac{2}{T^2} = \frac{2K}{T},
\end{aligned}$$

where in the last inequality, we used a sub-Gaussian tail bound in Lattimore & Szepesvári (2020):

Lemma F.1 (Corollary 5.5, Lattimore & Szepesvári (2020))**.** *Assume that $X_i - \mu$ are independent, σ-sub-Gaussian random variables. Then for any $\epsilon \ge 0$,*

$$\Pr(\hat{\mu} \ge \mu + \epsilon) \le \exp(-\frac{n\epsilon^2}{2\sigma^2}), \quad \Pr(\hat{\mu} \le \mu - \epsilon) \le \exp(-\frac{n\epsilon^2}{2\sigma^2})$$

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

Under the good event G, the optimal arm k^* is never eliminated since $\hat{\mu}_{k^*,T_{k^*}(t)} + 2\sqrt{\frac{\ln T}{T_{k^*}(t)}} \ge \mu^* > \mu_k \ge \hat{\mu}_{k,T_k(t)} - 2\sqrt{\frac{\ln T}{T_k(t)}}$, for all t, i. Furthermore, at time τ_k, by (12) we have

$$\mu^* - 4\sqrt{\frac{\ln T}{T_{k^*}(\tau_k)}} \le \mu_k + 4\sqrt{\frac{\ln T}{T_k(\tau_k)}}$$

$$\Delta_k \le 4\sqrt{\frac{\ln T}{T_{k^*}(\tau_k)}} + 4\sqrt{\frac{\ln T}{T_k(\tau_k)}}.$$

The FC-SE algorithm behaves in a strictly cyclic manner. An arm is pulled the same number of times in each feasible cycle. Since τ_k is at the end of a feasible cycle, we observe that

$$\frac{T_k(\tau_k)}{n_k} = \frac{T_{k^*}(\tau_k)}{n_{k^*}}. \tag{14}$$

Using this fact, we obtain

$$\Delta_k \le 4\sqrt{\frac{\ln T}{\frac{n_{k^*}}{n_k}T_k(\tau_k)}} + 4\sqrt{\frac{\ln T}{T_k(\tau_k)}} = 4\sqrt{\frac{\ln T}{T_k(\tau_k)}}\left(1 + \sqrt{\frac{n_k}{n_{k^*}}}\right),$$

which directly implies

$$T_k(\tau_k) \le \frac{16\ln T}{\Delta_k^2}\left(1 + \sqrt{\frac{n_k}{n_{k^*}}}\right)^2 \le \frac{32\ln T}{\Delta_k^2}\left(1 + \frac{n_k}{n_{k^*}}\right)$$

$$T_k(T) \le n_k + \frac{32\ln T}{\Delta_k^2}\left(1 + \frac{n_k}{n_{k^*}}\right).$$

The expected regret of FC-SE is then upper bounded as

$$\begin{aligned}
R_T =&\mathbb{E}\big[\sum_{t=1}^{T}\Delta_{a_t}\big] = \mathbb{E}\big[\sum_{k\neq k^*}^{K}\Delta_k T_k(T)\big]\\
=&\mathbb{E}\big[\sum_{k\neq k^*}^{K}\Delta_k T_k(T)[\mathbb{I}(G)+\mathbb{I}(\neg G)]\big]\\
\leq&\sum_{k\neq k^*}^{K}\Delta_k\mathbb{E}\big[T_k(T)\mathbb{I}(G)\big]+\Delta_{\max}T\Pr(\neg G)\\
\leq&\sum_{k\neq k^*}^{K}\left[\Delta_k n_k+\frac{32\ln T}{\Delta_k}\Big(1+\frac{n_k}{n_{k^*}}\Big)\right]+2K\Delta_{\max}.
\end{aligned}$$

□

Proof of Theorem 4.2. Let r_k be the index of cycle at the end of which arm k enters the feasible cycle. We say that the arms in the initial feasible cycle enter in the 0-th cycle, $\forall k:\mathrm{ind}(k)\leq N$, $r_k=0$. Since in the $N<K$ case, there are some arms entering the feasible cycle later in the game, the fact (14) in the proof of Theorem 4.1 no longer holds. Say t is the time that a feasible cycle ends. Arm k,k' have entered the feasible cycle before or at time t and not been eliminated before time t. Then we have

$$\frac{T_k(t)}{n_k}+r_k=\frac{T_{k'}(t)}{n_{k'}}+r_{k'}. \tag{15}$$

Following the analysis in the proof of Theorem 4.1, under the good event G (we use the definition in (13)), if arm j is not eliminated by the optimal active arm i^* at the end of some cycle t, we have

$$\Delta_{i^*,j}\leq 4\sqrt{\frac{\ln T}{T_j(t)}}+4\sqrt{\frac{\ln T}{T_{i^*}(t)}}.$$

Using (15), we obtain

$$\Delta_{i^*,j}\leq 4\sqrt{\frac{\ln T}{T_j(t)}}+4\sqrt{\frac{\ln T}{\frac{n_{i^*}}{n_j}T_j(t)+n_{i^*}(r_j-r_{i^*})}}.$$

We consider the following three different cases and get a uniform upper bound for $\frac{T_j(t)}{n_j}$.

(1) $r_j\geq r_{i^*}$. Arm j enters either after or simultaneously with the optimal active arm. We have

$$\Delta_{i^*,j}\leq 4\sqrt{\frac{\ln T}{T_j(t)}}+4\sqrt{\frac{\ln T}{\frac{n_{i^*}}{n_j}T_j(t)}}\leq 4\sqrt{\frac{\ln T}{T_j(t)}}\Big(1+\sqrt{\frac{n_j}{n_{i^*}}}\Big)$$

$$\frac{T_j(t)}{n_j}\leq\frac{16\ln T}{\Delta_{i^*,j}^2}\Big(1+\sqrt{\frac{n_j}{n_{i^*}}}\Big)^2\frac{1}{n_j}.$$

(2) $r_j<r_{i^*}$ and $T_j(t)>n_j(r_{i^*}-r_j)$. We have $T_j(t)>T_j(t)-n_j(r_{i^*}-r_j)$. Thus

$$\begin{aligned}
\Delta_{i^*,j}\leq&4\sqrt{\frac{\ln T}{T_j(t)-n_j(r_{i^*}-r_j)}}+4\sqrt{\frac{\ln T}{\frac{n_{i^*}}{n_j}T_j(t)+n_{i^*}(r_j-r_{i^*})}}\\
=&4\sqrt{\frac{\ln T}{T_j(t)-n_j(r_{i^*}-r_j)}}\Big(1+\sqrt{\frac{n_j}{n_{i^*}}}\Big)
\end{aligned}$$

$$\frac{T_j(t)}{n_j} \leq r_{i^*} - r_j + \frac{16 \ln T}{\Delta_{i^*,j}^2}\Big(1 + \sqrt{\frac{n_j}{n_{i^*}}}\Big)^2 \frac{1}{n_j}.$$

(3) $r_j < r_{i^*}$ and $T_j(t) \leq n_j(r_{i^*} - r_j)$. This directly implies

$$\frac{T_j(t)}{n_j} \leq r_{i^*} - r_j.$$

In conclusion, we have

$$\frac{T_j(t)}{n_j} \leq \max(0, r_{i^*} - r_j) + \frac{16 \ln T}{\Delta_{i^*,j}^2}\Big(1 + \sqrt{\frac{n_j}{n_{i^*}}}\Big)^2 \frac{1}{n_j}, \tag{16}$$

if arm j is not eliminated by i^* at t under the good event G.

Due to the patience tightening process in FC-SE, arm $\text{ind}^{-1}(N+a)$ is assigned to checkpoint $\lceil \frac{a}{N-1} \rceil c$, for $a \in [K-N]$. Let $\mathcal{C}$ denote the set of all checkpoints, then

$$\mathcal{C} = \left\{ \Big\lceil \frac{a}{N-1} \Big\rceil c \;\middle|\; a = 1, ..., K-N \right\} = \{c_1, c_2, ..., c_{|\mathcal{C}|}\},$$

where $c_1, c_2, ...$ is the ordered sequence of checkpoints: $c_1 < c_2 < ... < c_{|\mathcal{C}|}$. Now we go through the whole process from the beginning. Let $A_0 := \{k \in [K] : \text{ind}(k) \leq N\}$ be the arms in the initial cycle. Define $k^*(0) = \arg\max_{k \in A_0} \mu_k$. Consider the first time t_1 that triggers the checkpoint $c_1 = c$. It satisfies $t_1 + \sum_{k \in S} > c_1 - n$, where S is the set of arms in the feasible cycle after the arm elimination at time $t_1 - 1$. Let R_1 denote the number of the operated feasible cycles from time slot 1 to time slot $t_1 - 1$. $t_1 + \sum_{k \in S} > c_1 - n$ if and only if the length of R_1 feasible cycles plus the length of the feasible cycle beginning at time t_1 is at least $c_1 - n$. The length of all feasible cycles is upper bounded by n, thus we have $n(R_1 + 1) \geq c_1 - n$, $R_1 \geq \frac{c_1 - n}{n} - 1$. This inspires us to define

$$E_0 = \left\{ i \in A_0 \;\middle|\; i \neq k^*(0), \frac{16 \ln T}{\Delta_{k^*(0),i}^2}\Big(\frac{1}{\sqrt{n_i}} + 1\Big)^2 + 1 \leq \frac{c_1 - n}{n} - 1 \right\}. \tag{17}$$

For any $k \in E_0$, under the good event G defined in (13), must have been eliminated either before or at the arm elimination phase of the $\left(\left\lfloor \frac{16 \ln T}{\Delta_{k^*(0),k}^2}\Big(\frac{1}{\sqrt{n_k}} + 1\Big)^2 \right\rfloor + 1\right)$-th feasible cycle by (16). Since

$$\left\lfloor \frac{16 \ln T}{\Delta_{k^*(0),k}^2}\Big(\frac{1}{\sqrt{n_k}} + 1\Big)^2 \right\rfloor + 1 \leq \frac{16 \ln T}{\Delta_{k^*(0),k}^2}\Big(\frac{1}{\sqrt{n_k}} + 1\Big)^2 + 1 \leq \frac{c_1 - n}{n} - 1 \leq R_1,$$

arm k is eliminated before the condition $t + \sum_{k \in S} > c_1 - n$ is triggered. Thus for any arm $k' \in S$ when the condition is triggered, we have that $k' \notin E_0$. If $k' = k^*(0)$, $\Delta_{k^*(0),k'} = 0$. Otherwise,

$$\frac{16 \ln T}{\Delta_{k^*(0),k'}^2}\Big(\frac{1}{\sqrt{n_{k'}}} + 1\Big)^2 + 1 > \frac{c_1 - n}{n} - 1, \quad \Delta_{k^*(0),k'} < 4\sqrt{\frac{\ln T}{\frac{c_1 - n}{n} - 2}}\Big(\frac{1}{\sqrt{n_{k'}}} + 1\Big) = 4\sqrt{\frac{\ln T}{\frac{c_1}{n} - 3}}\Big(\frac{1}{\sqrt{n_{k'}}} + 1\Big).$$

Since it is possible that $k^*(0)$ is unfortunately dropped at checkpoint c_1, we can only guarantee that under G there is an arm in S with a mean reward at least $\mu_{k^*(0)} - 8\sqrt{\frac{\ln T}{\frac{c_1}{n} - 3}}$ after the operation at checkpoint c_1. The operation includes randomly dropping arms, adding new arms and rebuilding the feasible cycle. Define $k^*(1)$ as the temporarily optimal arm in the feasible cycle after the operation at checkpoint c_1. Then we have under the good event G,

$$\mu_{k^*(1)} \geq \max\left\{ \underline{\mu}_{k^*(0)} - 8\sqrt{\frac{\ln T}{\frac{c_1}{n} - 3}}, \max_{a \in [K]: m'_a = c_1} \mu_a \right\} =: \underline{\mu}_{k^*(1)}$$

with $\underline{\mu}_{k^*(0)} = \mu_{k^*(0)}$. Define $A_1 = (A_0 - E_0) \cup \{a \in [K] : m'_a = c_1\}$ as the set of arms possibly in the feasible cycle after the operation at checkpoint c_1.

Now we consider the situation after the operation at checkpoint c_{j-1} where $1 < j \leq 1 + |\mathcal{C}|$. Note that $c_{1+|\mathcal{C}|}$ is actually not a checkpoint in $\mathcal{C}$, but we define it as the time slot $c(1 + |\mathcal{C}|)$ for the convenience of our analysis. There is no any special operation at this "checkpoint" but the condition $t + \sum_{k\in S} n_k > c_{1+|\mathcal{C}|} - n$ is defined. Define R_{j-1} as the number of the operated feasible cycles from time slot 1 to time slot t_{j-1}, where t_{j-1} is the first time that triggers the checkpoint c_{j-1}. t_j, R_j are defined accordingly. Following the discussion concerning checkpoint c_1, we have that the length of the first R_{j-1} feasible cycles plus the length of the next feasible cycle is at least $c_{j-1} - n$, but the length of the first R_{j-1} feasible cycles is strictly less than $c_{j-1} - n$. Thus we observe that at least the time slots from $c_{j-1} - n$ to $c_j - n$ ($c_j - c_{j-1} + 1$ time slots in total) is covered by the $R_j - R_{j-1} + 1$ more feasible cycles after the operation at checkpoint c_{j-1}. We have $n(R_j - R_{j-1} + 1) \geq c_j - c_{j-1} + 1$, $R_j - R_{j-1} \geq \frac{c_j - c_{j-1}}{n} - 1$. This also inspires us to define

$$E_{j-1} = \left\{ i \in A_{j-1} \,\middle|\, \underline{\mu}_{k^*(j-1)} > \mu_i, \frac{16\ln T}{(\underline{\mu}_{k^*(j-1)} - \mu_i)^2}\Big(\frac{1}{\sqrt{n_i}} + 1\Big)^2 + 1 \leq \frac{c_j - c_{j-1}}{n} - 1 \right\},$$

where $A_{j-1} = (A_{j-2} - E_{j-2}) \cup \{a \in [K] : m'_a = c_{j-1}\}$ is the set of arms possibly in the feasible cycle after the operation at checkpoint c_{j-1} and $\underline{\mu}_{k^*(j-1)} := \max\big\{\underline{\mu}_{k^*(j-2)} - 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}, \max_{a\in[K]:m'_a = c_{j-1}} \mu_a\big\}$. For any $k \in E_{j-1}$, we will show that it must be eliminated before the operation at checkpoint c_j. Assume that arm k has not yet been eliminated after the operation at checkpoint c_{j-1}. By (16), under the good event G, arm k must have been eliminated either before or at the feasible cycle indexed by

$$r_k + \max(0, r_{k^*(j-1)} - r_k) + \left\lfloor \frac{16\ln T}{\Delta^2_{k^*(j-1),k}}\Big(\frac{1}{\sqrt{n_k}} + 1\Big)^2 \right\rfloor + 1.$$

Before the condition $t + \sum_{k\in S} > c_j - n$ at checkpoint c_j is triggered, the last feasible cycle where there can be new entrance is the R_{j-1}-th cycle. Thus $r_k, r_{k^*(j-1)} \leq R_{j-1}$. By induction hypothesis, $\mu_{k^*(j-1)} \geq \underline{\mu}_{k^*(j-1)}$. We have

$$\begin{aligned}
&r_k + \max(0, r_{k^*(j-1)} - r_k) + \left\lfloor \frac{16\ln T}{\Delta^2_{k^*(j-1),k}}\Big(\frac{1}{\sqrt{n_k}} + 1\Big)^2 \right\rfloor + 1\\
\leq& \max(r_{k^*(j-1)}, r_k) + \frac{16\ln T}{\Delta^2_{k^*(j-1),k}}\Big(\frac{1}{\sqrt{n_k}} + 1\Big)^2 + 1\\
\leq& R_{j-1} + \frac{16\ln T}{(\underline{\mu}_{k^*(j-1)} - \mu_k)^2}\Big(\frac{1}{\sqrt{n_k}} + 1\Big)^2 + 1\\
\leq& R_{j-1} + \frac{c_j - c_{j-1}}{n} - 1 \leq R_{j-1} + R_j - R_{j-1} = R_j.
\end{aligned}$$

So arm k must be eliminated before the operation at checkpoint c_j. Thus for any arm k' still in the feasible cycle when the condition at checkpoint c_j is triggered, we have that $k' \notin E_{j-1}$. We have either $\underline{\mu}_{k^*(j-1)} \leq \mu_{k'}$ or

$$\underline{\mu}_{k^*(j-1)} - \mu_{k'} < \sqrt{\frac{16\ln T}{\frac{c}{n} - 2}}\Big(\frac{1}{\sqrt{n_{k'}}} + 1\Big) \leq 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}.$$

Under the good event G, there must be an arm in the feasible cycle after the operation at checkpoint c_j whose mean reward is at least $\underline{\mu}_{k^*(j-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}$. Thus we have

$$\mu_{k^*(j)} \geq \max\left\{ \underline{\mu}_{k^*(j-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}, \max_{a\in[K]:m'_a = c_j} \mu_a \right\} =: \underline{\mu}_{k^*(j)}$$

under the good event G. Before formally deriving the expected regret upper bound for the FC-SE algorithm, we need to find a lower bound for $\underline{\mu}_{k^*(|\mathcal{C}|)}$ which indicates that randomly dropping arms does not severely decrease the highest mean reward in the active arm set. We consider two cases:

(1) $\mu^* = \max_{k \in A_0} \mu_k$. The optimal arm k^* is in the initial feasible cycle. $\mu^* = \mu_{k^*(0)} = \underline{\mu}_{k^*(0)}$. By the definition of $\underline{\mu}_{k^*(j)}, j = 0, 1, ..., |\mathcal{C}|$, we have

$$\underline{\mu}_{k^*(|\mathcal{C}|)} \geq \underline{\mu}_{k^*(|\mathcal{C}|-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}} \geq \underline{\mu}_{k^*(|\mathcal{C}|-2)} - 2 \times 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}} \geq ... \geq \underline{\mu}_{k^*(0)} - 8|\mathcal{C}|\sqrt{\frac{\ln T}{\frac{c}{n} - 3}} = \mu^* - 8|\mathcal{C}|\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}.$$

(2) $\exists j \in \{1, 2, ..., |\mathcal{C}|\}$, $\mu^* = \max_{a \in [K]: m'_a = c_j} \mu_a$. Say the optimal arm enters the feasible cycle at checkpoint c_{j^*}, thus

$$\underline{\mu}_{k^*(j^*)} = \max\left\{\underline{\mu}_{k^*(j^*-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}, \max_{a \in [K]: m'_a = c_{j^*}} \mu_a\right\} = \max\left\{\underline{\mu}_{k^*(j^*-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}, \mu^*\right\} = \mu^*.$$

As in the last case, we also obtain

$$\begin{aligned}\underline{\mu}_{k^*(|\mathcal{C}|)} &\geq \underline{\mu}_{k^*(|\mathcal{C}|-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}} \geq ... \\ &\geq \underline{\mu}_{k^*(j^*)} - 8(|\mathcal{C}| - j^*)\sqrt{\frac{\ln T}{\frac{c}{n} - 3}} \geq \mu^* - 8|\mathcal{C}|\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}.\end{aligned}$$

Now we derive the expected regret upper bound for the FC-SE algorithm

$$\begin{aligned}R_T =&\mathbb{E}\big[\sum_{t=1}^{T} \Delta_{a_t}\big] = \mathbb{E}\big[\sum_{k \neq k^*}^{K} \Delta_k T_k(T)\big] \\ =&\mathbb{E}\big[\sum_{k \neq k^*}^{K} \Delta_k T_k(T)[\mathbb{I}(G) + \mathbb{I}(\neg G)]\big] \\ \leq& \sum_{k \neq k^*}^{K} \Delta_k \mathbb{E}\big[T_k(T)\mathbb{I}(G)\big] + \Delta_{\max} T \Pr(\neg G) \\ =& \sum_{k \neq k^*: k \in E_{|\mathcal{C}|}} \Delta_k \mathbb{E}\big[T_k(T)\mathbb{I}(G)\big] + \sum_{k \neq k^*: k \notin E_{|\mathcal{C}|}} \Delta_k \mathbb{E}\big[T_k(T)\mathbb{I}(G)\big] + \Delta_{\max} T \Pr(\neg G).\end{aligned}$$

Since all arms have once entered the feasible cycle either before or at checkpoint $c_{|\mathcal{C}|}$, any arm $k \in E_{|\mathcal{C}|}$ must be eliminated before (virtual) checkpoint $c_{|\mathcal{C}|+1}$ under the good event G. Thus arm k is pulled at most $c(1 + |\mathcal{C}|)$ times. For arm $k' \notin E_{|\mathcal{C}|}$, $\mu_{k'}$ is not too small (i.e. $\mu_{k'} \geq \underline{\mu}_{k^*(|\mathcal{C}|)} - 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}$).

$$\begin{aligned}R_T \leq& \sum_{k \neq k^*: k \in E_{|\mathcal{C}|}} \Delta_k c(1 + |\mathcal{C}|) + \sum_{k \neq k^*: k \notin E_{|\mathcal{C}|}} (\mu^* - \underline{\mu}_{k^*(|\mathcal{C}|)} + \underline{\mu}_{k^*(|\mathcal{C}|)} - \mu_k)\mathbb{E}\big[T_k(T)\mathbb{I}(G)\big] + 2K\Delta_{\max} \\ \leq& \sum_{k \neq k^*: k \in E_{|\mathcal{C}|}} \Delta_k c(1 + |\mathcal{C}|) + \sum_{k \neq k^*: k \notin E_{|\mathcal{C}|}} 8(1 + |\mathcal{C}|)\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}\mathbb{E}\big[T_k(T)\big] + 2K\Delta_{\max} \\ \leq& \sum_{k \neq k^*} \Delta_k c(1 + |\mathcal{C}|) + 8(1 + |\mathcal{C}|)\sqrt{\frac{\ln T}{\frac{c}{n} - 3}} T + 2K\Delta_{\max} \\ \leq&(K - 1)\bar{\Delta} c(1 + |\mathcal{C}|) + 8T(1 + |\mathcal{C}|)\sqrt{\frac{n \ln T}{c - 3n}} + 2K\Delta_{\max}.\end{aligned}$$

The last formula is minimized when

$$\begin{aligned}c &= 3n + \left(\frac{8T(1 + |\mathcal{C}|)\sqrt{n \ln T}}{2(K - 1)\bar{\Delta}(1 + |\mathcal{C}|)}\right)^{\frac{2}{3}} \\ &= 3n + \left(\frac{4T\sqrt{n \ln T}}{(K - 1)\bar{\Delta}}\right)^{\frac{2}{3}}.\end{aligned}$$

By the definition of c in (4), if $\left\lfloor \min_{k:\text{ind}(k)>N} \frac{m_k}{\lceil \frac{\text{ind}(k)-N}{N-1} \rceil} \right\rfloor > 3n + \left\lceil \left(\frac{4T\sqrt{n \ln T}}{(K-1)\bar{\Delta}} \right)^{\frac{2}{3}} \right\rceil$, we have

$$\begin{aligned} R_T \leq & (K-1)\bar{\Delta}\left(3n + \left\lceil \left(\frac{4T\sqrt{n \ln T}}{(K-1)\bar{\Delta}} \right)^{\frac{2}{3}} \right\rceil\right)\left(1 + \left\lceil \frac{K-N}{N-1} \right\rceil\right) \\ & + 8T\left(1 + \left\lceil \frac{K-N}{N-1} \right\rceil\right)\sqrt{\frac{n \ln T}{\left(\frac{4T\sqrt{n \ln T}}{(K-1)\bar{\Delta}} \right)^{\frac{2}{3}}}} + 2K\Delta_{\max} \\ \leq & (K-1)\bar{\Delta}\left(1 + 3n + \left(\frac{4T\sqrt{n \ln T}}{(K-1)\bar{\Delta}} \right)^{\frac{2}{3}}\right)\left(1 + \left\lceil \frac{K-N}{N-1} \right\rceil\right) \\ & + 8\left(1 + \left\lceil \frac{K-N}{N-1} \right\rceil\right)4^{-\frac{1}{3}}T^{\frac{2}{3}}(n \ln T)^{\frac{1}{3}}(K-1)^{\frac{1}{3}}\bar{\Delta}^{\frac{1}{3}} + 2K\Delta_{\max}. \end{aligned}$$

Thus $R_T = O\big(K^{\frac{4}{3}}T^{\frac{2}{3}}(n \ln T)^{\frac{1}{3}}\big)$.

□

G. Details of FC-Entry Algorithm

The details of Feasible Cycle-based successive elimination with new Entering arms (FC-Entry) is given in Algorithm 4. For notational simplicity, $\boldsymbol{m}$ is written to be one of the inputs for Algorithm 4, but it is important to note that the algorithm only needs the patience m_k of the initially available arms (i.e. $k \in [K_0]$). There is a similar patience tightening operation for the arms initially available but not included in the initial feasible cycle, as in Algorithm 1. We still define checkpoints as integer multiples of a constant c, while the c for FC-Entry algorithm is instead given in (11). S is the set contain all arms actually in the feasible cycle, while S_{res} is the set of arms occupying the reserved slots. When arm k enters the game at time ρ_k, the algorithm immediately adds it to the feasible cycle (it is possible that the algorithm randomly drops some arm in the reserved places to make room for arm k). Thus the algorithm adds arm k to S and S_{res}. However, the algorithm adds arm k to the feasible cycle only to prevent its departure. For the ease of theoretical analysis, we require that the rewards generated by a newly entered arm k before it is 'formally' involved in the feasible cycle are not used to compute its estimated reward mean. A new entering arm is formally involved in the feasible cycle only in the operation at the nearest checkpoint after its entrance. It does not either participate in the arm elimination phase before that checkpoint. When a new entering arm is detected, the algorithm computes the next checkpoint if necessary. Before formally involved in the feasible cycle, the newly entered arm temporarily stays in S_{pre}. For any arm $k \in [K]$, $\xi_k, T_k^{\flat}$ record the sum of observed rewards and the number of pulls after arm k is formally involved in the feasible cycle, respectively. At the end of each feasible cycle, the algorithm checks whether the p-th checkpoint is reached. Checkpoints $c_1, ..., c_{\lceil \frac{K_0-N+2}{N-3} \rceil}$ are known in advance since at these checkpoints the algorithm involves the arms in $\{k \in [K_0] : \text{ind}(k) > N-2\}$ into the feasible cycle. However, if after the operation at checkpoint $c_{\lceil \frac{K_0-N+2}{N-3} \rceil}$ there is still new entering arms that have not formally enter the feasible cycle, the algorithm needs to compute the previously unknown c_p for $p > \lceil \frac{K_0-N+2}{N-3} \rceil$ when detecting new entering arms. Since t denotes the first time slot of the next feasible cycle, $\bar{t} := t + \sum_{k \in S-S_{\text{res}}} n_k + n_+ + n_-$ is an upper bound for the first time slot of the cycle after the next feasible cycle. Let t' denote the first time slot of the last pulled feasible cycle. The condition is not triggered then: $t \leq t' + \sum_{k \in S'-S'_{\text{res}}} n_k + n_+ + n_- \leq c_p - n$, where S', S'_{res} are the versions of S, S_{res} when checking the condition at time $t'-1$. S_{new} is not empty, we have $p \leq \lceil \frac{K_0-N+2}{N-3} \rceil$. Each arm k in S_{new} is pulled no later than $t + n - 1 \leq c_p - 1 < c_p \leq m'_k \leq m_k$. Thus arm k does not leave early. In Theorem 4.6 we present an expected dynamic regret upper bound for FC-Entry algorithm under the sparse entrance assumption: $3c \leq \min_{K_0<k<K}(\rho_{k+1} - \rho_k - 1)$.

Proof of Theorem 4.6. The set of all check points is

$$\mathcal{C} = \left\{ \left\lceil \frac{a}{N-3} \right\rceil c \;\middle|\; a = 1, ..., K_0 - N + 2 \right\} \cup \{ j_k c \mid k > K_0 \} = \{c_1, c_2, ..., c_{|\mathcal{C}|}\},$$

where j_k is defined such that arm k is involved in the feasible cycle at the j_k-th checkpoint. Specifically, $j_k = 0$ for $k\ s.t.\ k \in [K_0]$ and $\text{ind}(k) \leq N-2$, while $j_k = \lceil \frac{\text{ind}(k)-N+2}{N-3} \rceil$ for $k\ s.t.\ k \in [K_0]$ and $\text{ind}(k) > N-2$. j_k for $k > K_0$ is

Algorithm 4 FC-Entry

```
**Input:** Number of arms in the initial feasible cycle $N$, patience vector $\boldsymbol{m}$, time horizon $T$, a lower bound for the new
  entering arms' patience $\underline{m}$, segment length $c$
Construct a feasible cycle "$\bar{a}_1, ..., \bar{a}_n$" for the set of arms $\{k : \text{ind}(k) \leq N-2\} \cup \{+,-\}$
$S \leftarrow \{k \in [K_0] : \text{ind}(k) \leq N-2\}, S_{\text{pre}} \leftarrow \emptyset, S_{\text{res}} \leftarrow \emptyset, p \leftarrow 1, e \leftarrow K_0+1,$ "$\bar{a}'_1, ..., \bar{a}'_n$" $\leftarrow$ "$\bar{a}_1, ..., \bar{a}_n$"
$\bar{a}_i \leftarrow 0, \forall i : \bar{a}'_i \in \{+,-\}$. $\xi_k \leftarrow 0, T^\flat_k \leftarrow 0, \forall k \in [K]$. $t \leftarrow 1, \bar{t} \leftarrow t + \sum_{k \in S-S_{\text{res}}} n_k + n_+ + n_-$
for $k \in \{k' \in [K_0] : \text{ind}(k') > N-2\}$ do
    $m'_k \leftarrow \left\lceil \frac{\text{ind}(k)-N+2}{N-3} \right\rceil c$
end for
while $t \leq T$ do
    for $i = 1, ..., n$ do
        if $e \leq K$ and $t = \rho_e$ then
            If $|S_{\text{res}}| = 2$ randomly choose $* \in \{+,-\}$, $S \leftarrow S - \{a\}$, $S_{\text{res}} \leftarrow S_{\text{res}} - \{a\}$ such that $\bar{a}_i = a, \forall i : \bar{a}'_i = *$.
              $\bar{a}_i \leftarrow 0$ for $i = 1, ..., n$ s.t. $\bar{a}_i \notin S$
            Randomly choose $* \in \{+,-\}$ $s.t.$ $\bar{a}_i = 0, \forall i : \bar{a}'_i = *$, $S \leftarrow S \cup \{e\}$, $S_{\text{res}} \leftarrow S_{\text{res}} \cup \{e\}$, $S_{\text{pre}} \leftarrow S_{\text{pre}} \cup \{e\}$,
              $\bar{a}_i \leftarrow e$ for $i = 1, ..., n$ s.t. $\bar{a}'_i = *$
            if $p > \lceil \frac{K_0-N+2}{N-3} \rceil$ then
                $c_p \leftarrow \min\{j'c \mid j' \in \mathbb{N}^+, \bar{t} \leq j'c - n\}$
            end if
            $e \leftarrow e + 1$
        end if
        if $\bar{a}_i \neq 0$ then
            Pull $a_t = \bar{a}_i$ and receive reward $X_{a_t, T_{a_t}(t)}$. $\xi_{a_t} \leftarrow \xi_{a_t} + X_{a_t, T_{a_t}(t)}$ and $T^\flat_{a_t} \leftarrow T^\flat_{a_t} + 1$ if $a_t \notin S_{\text{pre}}$
            $t \leftarrow t + 1$
        end if
    end for
    $S \leftarrow S_{\text{pre}} \cup \left\{ k \in S - S_{\text{pre}} : \forall j \in S - S_{\text{pre}}, \frac{\xi_j}{T^\flat_j} - 2\sqrt{\frac{\ln T}{1 \vee T^\flat_j}} \leq \frac{\xi_k}{T^\flat_k} + 2\sqrt{\frac{\ln T}{1 \vee T^\flat_k}} \right\}$
    $S_{\text{res}} \leftarrow S_{\text{res}} \cap S$
    $\bar{a}_i \leftarrow 0$ for $i = 1, ..., n$ s.t. $\bar{a}_i \notin S$
    $\bar{t} \leftarrow t + \sum_{k \in S-S_{\text{res}}} n_k + n_+ + n_-$
    if $c_p$ is known and $\bar{t} > c_p - n$ then
        $S_{\text{pre}} \leftarrow \emptyset$, $S_{\text{new}} \leftarrow \{k \in [K_0] : m'_k = c_p\}$
        while $|S - S_{\text{res}}| + |S_{\text{new}}| > N - 2$ do
            $a \sim \text{Unif}(S - S_{\text{res}})$
            $S \leftarrow S - \{a\}$
        end while
        $\bar{a}_i \leftarrow 0$ for $i = 1, ..., n$ s.t. $\bar{a}_i \notin S$
        while $|S_{\text{new}}| > 0$ do
            $a \leftarrow \arg\min_{k \in S_{\text{new}}} \text{ind}(k)$
            $S \leftarrow S \cup \{a\}, S_{\text{new}} \leftarrow S_{\text{new}} - \{a\}$
            $\bar{a}_i \leftarrow a$ where $i = \min\left\{i' \leq n : \bar{a}'_{i'} = \min\{k \in [K_0] \mid \text{ind}(k) \leq N-2 \text{ and } \forall j \leq n \; s.t. \; \bar{a}'_j = k : \bar{a}_j = 0\}\right\}$
        end while
        $p \leftarrow p + 1$
    end if
end while
```

a random variable though ρ_k is fixed because the feasible cycles can have various lengths. Besides, since we assume that ρ_k for $k > K_0$ is unknown in advance, only the first part of $\mathcal{C}$ (i.e. $\left\{ \left\lceil \frac{a}{N-3} \right\rceil c \,\middle|\, a = 1, ..., K_0 - N + 2 \right\}$) is known in the beginning of the game.

Let $A_0 = \{k \in [K_0] : \text{ind}(k) \leq N-2\}$ be the set of arms in the initial feasible cycle. Define $k^*(0) = \arg\max_{k \in A_0} \mu_k$. We define the same E_0 as in (17). We note that if there is a new entering arm before the condition at checkpoint c_1 is triggered, no arm in S_{res} is kicked off because in the beginning, the reserved positions in the feasible cycle are empty. Thus following the proof of Theorem 4.2, we also have that any arm $k \in E_0$ is eliminated before checkpoint c_1 under the good event G defined in (13), while any arm $k' \in S - S_{\text{pre}}$ when the condition $t + \sum_{k \in S - S_{\text{res}}} n_k + n_+ + n_- > c_p - n$ is triggered satisfies that

$$\Delta_{k^*(0),k'} < 4\sqrt{\frac{\ln T}{\frac{c_1 - n}{n} - 2}}\Big(\frac{1}{\sqrt{n_{k'}}} + 1\Big) = 4\sqrt{\frac{\ln T}{\frac{c_1}{n} - 3}}\Big(\frac{1}{\sqrt{n_{k'}}} + 1\Big).$$

Define $k^*(1)$ as the temporarily optimal arm in the feasible cycle after the operation at checkpoint c_1. Then we have under G,

$$\mu_{k^*(1)} \geq \max\left\{\underline{\mu}_{k^*(0)} - 8\sqrt{\frac{\ln T}{\frac{c_1}{n} - 3}}, \max_{a \in [K_0]: m'_a = c_1} \mu_a, \max_{a > K_0 : j_a = 1} \mu_a\right\} =: \underline{\mu}_{k^*(1)}$$

with $\underline{\mu}_{k^*(0)} = \mu_{k^*(0)}$. Define $A_1 := (A_0 - E_0) \cup \{a \in [K_0] : m'_a = c_1\} \cup \{a > K_0 : j_a = 1\}$ as the set of arms possibly in the feasible cycle after the operation at checkpoint c_1. Under the good event G, $\underline{\mu}_{k^*(1)}$ is a lower bound for the mean reward of the temporarily optimal arm in the feasible cycle after the operation at checkpoint c_1.

Now we consider the situation after the operation at checkpoint c_{j-1} where $1 < j \leq \lceil \frac{K_0 - N + 2}{N-3} \rceil$. The induction hypothesis is that under G, $\underline{\mu}_{k^*(j-1)}$ is a lower bound for the mean reward of the temporarily optimal arm in the feasible cycle after the operation at checkpoint c_{j-1}. Define R_{j-1} as the number of the operated feasible cycles from time slot 1 to time slot t_{j-1}, where t_{j-1} is the first time that triggers the checkpoint c_{j-1}. t_j, R_j are defined accordingly. When the condition at checkpoint c_{j-1} is triggered, we also have that $t'_{j-1} + \sum_{k \in S - S_{\text{res}}} n_k + n_+ + n_- \leq c_{j-1} - n$, where t'_{j-1} is the last time when arm elimination is performed before t_{j-1}. Since $\sum_{k \in S - S_{\text{res}}} n_k + n_+ + n_-$ is an upper bound for the length of the feasible cycle beginning at time t'_{j-1}, we have $t_{j-1} \leq t'_{j-1} + \sum_{k \in S - S_{\text{res}}} n_k + n_+ + n_- \leq c_{j-1} - n$. At least the time slots from $c_{j-1} - n$ to $c_j - n$ is covered by the $R_j - R_{j-1} + 1$ more feasible cycles after the operation at checkpoint c_{j-1}. We have $n(R_j - R_{j-1} + 1) \geq c_j - c_{j-1} + 1$, $R_j - R_{j-1} \geq \frac{c_j - c_{j-1}}{n} - 1$.

To compute $\underline{\mu}_{k^*(j)}$, we consider the following cases:

1. No entering arm k whose $j_k = j$. Under G, the temporarily optimal arm $k^*(j-1)$ has a reward mean at least $\underline{\mu}_{k^*(j-1)}$. Following the analysis in the proof of Theorem 4.2, any $k' \in S$ when the condition at checkpoint c_j is triggered satisfies that $\underline{\mu}_{k^*(j-1)} - \mu_{k'} \leq 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}$. We have $\underline{\mu}_{k^*(j)} \geq \underline{\mu}_{k^*(j-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}$.

2. A new entering arm k is involved in the feasible cycle at checkpoint c_j, but $|S_{\text{res}}| < 2$ when it enters the game. The new entering arm does not kick off any arm, thus the temporarily optimal arm with reward mean at least $\underline{\mu}_{k^*(j-1)}$ must be still active. We also have $\underline{\mu}_{k^*(j)} \geq \underline{\mu}_{k^*(j-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}$.

3. A new entering arm k is involved in the feasible cycle at checkpoint c_j and $|S_{\text{res}}| = 2$ when it enters the game, but the temporarily optimal arm is not kicked off. As long as the temporarily optimal arm is still active, we have $\underline{\mu}_{k^*(j)} \geq \underline{\mu}_{k^*(j-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}$.

4. A new entering arm k is involved in the feasible cycle at checkpoint c_j and it kicks off the temporarily optimal arm when it enters the game. Let $+, -$ represent the arms in S_{res} when arm k enters the game. In this case, by the sparse entrance assumption, $\max(j_+, j_-) \leq j - 2$. $|\mu_+ - \mu_-| \leq 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}$ because otherwise one of $\{+, -\}$ will be eliminated by the other either before or at checkpoint c_{j-1}. Thus although arm k kicks off the temporarily optimal arm, another arm in S_{res} has a mean reward at least $\underline{\mu}_{k^*(j-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n} - 3}}$. If this arm becomes temporarily optimal, it must be still in

the feasible cycle after the operation at checkpoint c_j, thus $\underline{\mu}_{k^*(j)} \geq \underline{\mu}_{k^*(j-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n}-3}}$. But if not, the temporarily optimal arm is in $S - S_{\text{res}}$ and has a reward mean at least $\underline{\mu}_{k^*(j-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n}-3}}$. Though this arm may be unfortunately dropped at checkpoint c_j, it guarantees that $\underline{\mu}_{k^*(j)} \geq \underline{\mu}_{k^*(j-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n}-3}} - 8\sqrt{\frac{\ln T}{\frac{c}{n}-3}} = \underline{\mu}_{k^*(j-1)} - 16\sqrt{\frac{\ln T}{\frac{c}{n}-3}}$.

Thus we define E_{j-1} as

$$\left\{ i \in A_{j-1} \,\middle|\, \underline{\mu}_{k^*(j-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n}-3}} > \mu_i, \frac{16\ln T}{\left(\underline{\mu}_{k^*(j-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n}-3}} - \mu_i\right)^2}\left(\frac{1}{\sqrt{n_i}}+1\right)^2 + 1 \leq \frac{c_j - c_{j-1}}{n} - 1 \right\},$$

where $A_{j-1} := (A_{j-2} - E_{j-2}) \cup \{a \in [K_0] : m'_a = c_{j-1}\} \cup \{a > K_0 : j_a = j-1\}$. For any arm $k \in E_{j-1}$, if it is in S after the operation at checkpoint c_{j-1}, we show that it will be eliminated before checkpoint c_j. By our discussion above, there is an arm v with reward mean at least $\underline{\mu}_{k^*(j-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n}-3}}$ which survives until the operation at checkpoint c_j. We have

$$\begin{aligned}
& r_k + \max(0, r_v - r_k) + \left\lfloor \frac{16\ln T}{\Delta_{v,k}^2}\left(\frac{1}{\sqrt{n_k}}+1\right)^2 \right\rfloor + 1 \\
=& \max(r_v, r_k) + \left\lfloor \frac{16\ln T}{\Delta_{v,k}^2}\left(\frac{1}{\sqrt{n_k}}+1\right)^2 \right\rfloor + 1 \\
\leq& \max(r_v, r_k) + \frac{16\ln T}{\Delta_{v,k}^2}\left(\frac{1}{\sqrt{n_k}}+1\right)^2 + 1 \\
\leq& R_{j-1} + \frac{16\ln T}{\left(\underline{\mu}_{k^*(j-1)} - 8\sqrt{\frac{\ln T}{\frac{c}{n}-3}} - \mu_k\right)^2}\left(\frac{1}{\sqrt{n_k}}+1\right)^2 + 1 \\
\leq& R_{j-1} + \frac{c_j - c_{j-1}}{n} - 1 \\
\leq& R_{j-1} + R_j - R_{j-1} \\
=& R_j.
\end{aligned}$$

Thus for any k' still in the feasible cycle when the condition at checkpoint c_j is triggered, we have $k' \notin E_{j-1}$ and

$$\mu_{k'} \geq \underline{\mu}_{k^*(j-1)} - 16\sqrt{\frac{\ln T}{\frac{c}{n}-3}}.$$

In conclusion, under the good event G,

$$\mu_{k^*(j)} \geq \max\left\{ \underline{\mu}_{k^*(j-1)} - 16\sqrt{\frac{\ln T}{\frac{c}{n}-3}}, \max_{a\in[K_0]:m'_a=c_j} \mu_a, \max_{a>K_0:j_a=j} \mu_a \right\} =: \underline{\mu}_{k^*(j)}.$$

A_{j-1}, E_{j-1} can be defined the same way for $j = \lceil \frac{K_0-N+2}{N-3} \rceil + 1, ..., |\mathcal{C}|+1$ by setting $c_{1+|\mathcal{C}|} = c_{|\mathcal{C}|} + c$. The difference is that $c_j - c_{j-1}$ can be larger than c and there is no arm dropped out at c_j. The temporarily optimal arm can still be kicked off by a new entering arm. With a similar discussion as in the $j \leq \lceil \frac{K_0-N+2}{N-3} \rceil$ case, we obtain

$$\mu_{k^*(j)} \geq \max\left\{ \underline{\mu}_{k^*(j-1)} - 16\sqrt{\frac{\ln T}{\frac{c}{n}-3}}, \max_{a>K_0:j_a=j} \mu_a \right\} =: \underline{\mu}_{k^*(j)}$$

for $j = \lceil \frac{K_0-N+2}{N-3} \rceil + 1, ..., |\mathcal{C}|$.

Whether the optimal arm k^* is in the initial feasible cycle is crucial for our regret analysis. We consider the following cases:

1. $k^* > K_0$. The optimal arm k^* enters later in the game. Let $\kappa_0^* = \arg\max_{k\in[K_0]} \mu_k$ be the optimal arm among those enter at the beginning of the game. Let $\kappa_j^* = \min\{k > K_0 : \mu_k > \mu_{\kappa^*_{j-1}}\}$ be the first entering arm whose mean reward is greater than $\mu_{\kappa^*_{j-1}}$. Integer τ is defined such that $\kappa^*_\tau = k^*$, we have $\tau > 0$ since $k^* > K_0$. By these definitions, we have

$$\mu_{\kappa_0^*} < \mu_{\kappa_1^*} < ... < \mu_{\kappa_\tau^*} = \mu_{k^*} = \mu^*.$$

And we can rewrite the expected dynamic regret $\tilde{R}_T$ as

$$\begin{aligned}
\tilde{R}_T =&\mathbb{E}\big[\sum_{t=1}^{T}\mu_t^* - \sum_{t=1}^{T}\mu_{a_t}\big] = \mathbb{E}\big[\sum_{t=1}^{\rho_{\kappa_1^*}-1}(\mu_{\kappa_0^*} - \mu_{a_t}) + \sum_{t=\rho_{\kappa_1^*}}^{\rho_{\kappa_2^*}-1}(\mu_{\kappa_1^*} - \mu_{a_t}) + ... + \sum_{t=\rho_{k^*}}^{T}(\mu^* - \mu_{a_t})\big]\\
=&\mathbb{E}\big[\sum_{t=1}^{\rho_{\kappa_1^*}-1}(\mu_{\kappa_0^*} - \mu_{a_t})(\mathbb{I}\{G\} + \mathbb{I}\{\neg G\}) + \sum_{t=\rho_{\kappa_1^*}}^{\rho_{\kappa_2^*}-1}(\mu_{\kappa_1^*} - \mu_{a_t})(\mathbb{I}\{G\} + \mathbb{I}\{\neg G\}) + ...\\
&+ \sum_{t=\rho_{k^*}}^{T}(\mu^* - \mu_{a_t})(\mathbb{I}\{G\} + \mathbb{I}\{\neg G\})\big]\\
\leq&\mathbb{E}\big[\sum_{t=1}^{\rho_{\kappa_1^*}-1}(\mu_{\kappa_0^*} - \mu_{a_t})\mathbb{I}\{G\} + \sum_{t=\rho_{\kappa_1^*}}^{\rho_{\kappa_2^*}-1}(\mu_{\kappa_1^*} - \mu_{a_t})\mathbb{I}\{G\} + ... + \sum_{t=\rho_{k^*}}^{T}(\mu^* - \mu_{a_t})\mathbb{I}\{G\} + \sum_{t=1}^{T}\Delta_{a_t}\mathbb{I}\{\neg G\}\big]\\
\leq&T\Delta_{\max}\Pr(\neg G) + \sum_{k\neq k^*}^{K}\mathbb{E}\big[\sum_{t=1}^{\rho_{\kappa_1^*}-1}(\mu_{\kappa_0^*} - \mu_k)\mathbb{I}\{a_t = k\}\mathbb{I}\{G\}\\
&+ \sum_{t=\rho_{\kappa_1^*}}^{\rho_{\kappa_2^*}-1}(\mu_{\kappa_1^*} - \mu_k)\mathbb{I}\{a_t = k\}\mathbb{I}\{G\} + ... + \sum_{t=\rho_{k^*}}^{T}(\mu^* - \mu_k)\mathbb{I}\{a_t = k\}\mathbb{I}\{G\}\big]\\
=:&T\Delta_{\max}\Pr(\neg G) + \sum_{k\neq k^*}^{K}\mathbb{E}[D_k].
\end{aligned}$$

To upper bound D_k, we define

$$l(k) = \begin{cases} \min\Big\{l = 0, ..., \tau\Big|\mu_{\kappa_l^*} - 16(K-N+3)\sqrt{\frac{\ln T}{\frac{c}{n}-3}} > \mu_k\Big\}, & \text{if } \Delta_k > 16(K-N+3)\sqrt{\frac{\ln T}{\frac{c}{n}-3}}.\\ \tau + 1, & \text{if } \Delta_k \leq 16(K-N+3)\sqrt{\frac{\ln T}{\frac{c}{n}-3}}.\end{cases}$$

and consider three possibilities:

(a) $l(k) = \tau + 1$. In this case, μ_k is very close to μ^*, thus we trivially bound $D_k \leq \Delta_k T \leq 16(K-N+3)T\sqrt{\frac{\ln T}{\frac{c}{n}-3}}$.

(b) $l(k) \leq \tau, j_k \geq j_{\kappa^*_{l(k)}}$. At the $j_{\kappa^*_{l(k)}}$-th checkpoint, arm $\kappa^*_{l(k)}$ enters the feasible cycle. Among those have once entered the feasible cycle, it has the highest mean reward. Thus by our previous definition, $\underline{\mu}_{k^*(j_{\kappa^*_{l(k)}})} = \mu_{\kappa^*_{l(k)}}$. The number of checkpoints is $|\mathcal{C}| \leq \lceil\frac{K_0-N+2}{N-3}\rceil + K - K_0 \leq K_0 - N + 2 + K - K_0 = K - N + 2$. Since arm k enters the feasible cycle either after or simultaneously with arm $\kappa^*_{l(k)}$ (i.e. $j_k \geq j_{\kappa^*_{l(k)}}$), there must be an arm with reward mean at least $\mu_{\kappa^*_{l(k)}} - 16|\mathcal{C}|\sqrt{\frac{\ln T}{\frac{c}{n}-3}} - 8\sqrt{\frac{\ln T}{\frac{c}{n}-3}}$ after the operation at checkpoint c_{j_k} which survives until at least the next checkpoint. Given that $\mu_{\kappa^*_{l(k)}} - 16(K-N+2)\sqrt{\frac{\ln T}{\frac{c}{n}-3}} - 8\sqrt{\frac{\ln T}{\frac{c}{n}-3}} - \mu_k > 8\sqrt{\frac{\ln T}{\frac{c}{n}-3}}$, arm k is eliminated before time slot $c_{j_k} + c - n$ under the good event G. Thus the total number of pulls of arm k is at most $3c$ and $D_k \leq 3c\Delta_k$.

(c) $l(k) \leq \tau, j_k < j_{\kappa^*_{l(k)}}$. Arm $\kappa^*_{l(k)}$ enters the game at time $\rho_{\kappa^*_{l(k)}}$. For notational simplicity, let c_κ denote the checkpoint at which arm $\kappa^*_{l(k)}$ is involved in the feasible cycle. We show that $c_\kappa - c - 2n \leq \rho_{\kappa^*_{l(k)}}$. Suppose this inequality does not hold. Let t' be the first time slot of the feasible cycle that contains the time slot $\rho_{\kappa^*_{l(k)}}$, $t' \leq \rho_{\kappa^*_{l(k)}}$. Thus

$t' + n \leq \rho_{\kappa^*_{l(k)}} + n < c_\kappa - c - n$. Arm $\kappa^*_{l(k)}$ should be involved in the feasible cycle at checkpoint $c_\kappa - c$ instead of c_κ, which is a contradiction. If arm k is still active when the condition at checkpoint c_κ is triggered, it must be eliminated before time slot $c_\kappa + c - n$ under the good event G since $\mu_{\kappa^*_{l(k)}} - \mu_k > 16(K-N+3)\sqrt{\frac{\ln T}{\frac{c}{n}-3}}$. Then we have $\sum_{t=\rho_{\kappa^*_{l(k)}}}^{\rho_{\kappa^*_{l(k)+1}}-1} \mathbb{I}\{a_t = k\}\mathbb{I}\{G\} \leq c_\kappa + c - n - \rho_{\kappa^*_{l(k)}} \leq c_\kappa + c - n - c_\kappa + c + 2n \leq 3c$ where we set $\rho_{\kappa^*_{\tau+1}} = T+1$. We derive an upper bound for D_k,

$$\begin{aligned}
D_k = &\sum_{t=1}^{\rho_{\kappa^*_1}-1} (\mu_{\kappa^*_0} - \mu_k)\mathbb{I}\{a_t = k\}\mathbb{I}\{G\} + \sum_{t=\rho_{\kappa^*_1}}^{\rho_{\kappa^*_2}-1} (\mu_{\kappa^*_1} - \mu_k)\mathbb{I}\{a_t = k\}\mathbb{I}\{G\} + ... + \sum_{t=\rho_{k^*}}^{T} (\mu^* - \mu_k)\mathbb{I}\{a_t = k\}\mathbb{I}\{G\} \\
\leq &\sum_{t=1}^{\rho_{\kappa^*_1}-1} (\mu_{\kappa^*_0} - \mu_k) + \sum_{t=\rho_{\kappa^*_1}}^{\rho_{\kappa^*_2}-1} (\mu_{\kappa^*_1} - \mu_k) + ... + \Delta_k \sum_{t=\rho_{\kappa^*_{l(k)}}}^{\rho_{\kappa^*_{l(k)+1}}-1} \mathbb{I}\{a_t = k\}\mathbb{I}\{G\} \\
\leq &(\rho_{\kappa^*_{l(k)}} - 1)16(K-N+3)\sqrt{\frac{\ln T}{\frac{c}{n}-3}} + 3c\Delta_k \leq 16(K-N+3)T\sqrt{\frac{\ln T}{\frac{c}{n}-3}} + 3c\Delta_k.
\end{aligned}$$

Then $\tilde{R}_T$ can be further bounded as

$$\tilde{R}_T \leq 2K\Delta_{\max} + \sum_{k \neq k^*}^{K} \mathbb{E}[D_k] \leq 2K\Delta_{\max} + \sum_{k \neq k^*}^{K} \left(16(K-N+3)T\sqrt{\frac{\ln T}{\frac{c}{n}-3}} + 3c\Delta_k\right).$$

2. $k^* \leq K_0, \text{ind}(k^*) > N-2$. The optimal arm k^* is in the game from the beginning.

$$\tilde{R}_T = \mathbb{E}\big[\sum_{t=1}^{T} \mu^*_t - \sum_{t=1}^{T} \mu_{a_t}\big] \leq T\Delta_{\max}\Pr(\neg G) + \sum_{k \neq k^*}^{K} \mathbb{E}[D_k],$$

where in this case $D_k = \Delta_k \sum_{t=1}^{T} \mathbb{I}\{a_t = k\}\mathbb{I}\{G\} = \Delta_k T_k(T)\mathbb{I}\{G\}$. Arm k^* is involved in the feasible cycle at checkpoint $c_{\lceil \frac{\text{ind}(k^*)-N+2}{N-3} \rceil}$. Consider the case when $\Delta_k > 16(K-N+3)\sqrt{\frac{\ln T}{\frac{c}{n}-3}}$. By our previous discussion, it guarantees that $T_k(T) \leq 3c$ when $j_k \geq j_{k^*}$ under the good event G. When $j_k < j_{k^*}$, arm k must be eliminated before time slot $c_{k^*} + c - n$. Thus $\Delta_k > 16(K-N+3)\sqrt{\frac{\ln T}{\frac{c}{n}-3}}$ implies that $\mathbb{I}\{G\}T_k(T) \leq 3c + c_{k^*} \leq 3c + c\lceil \frac{K_0-N+2}{N-3} \rceil$. Since it is also possible that $\Delta_k \leq 16(K-N+3)\sqrt{\frac{\ln T}{\frac{c}{n}-3}}$, we have $D_k \leq 3c\Delta_k + \Delta_k c\lceil \frac{K_0-N+2}{N-3} \rceil + 16(K-N+3)T_k(T)\sqrt{\frac{\ln T}{\frac{c}{n}-3}}$. Thus

$$\tilde{R}_T \leq 2K\Delta_{\max} + \sum_{k \neq k^*}^{K} \left(3c\Delta_k + \Delta_k c\left\lceil \frac{K_0-N+2}{N-3} \right\rceil + 16(K-N+3)T_k(T)\sqrt{\frac{\ln T}{\frac{c}{n}-3}}\right).$$

3. $k^* \leq K_0, \text{ind}(k^*) \leq N-2$. The optimal arm k^* is in the initial feasible cycle. The only difference from the $k^* \leq K_0, \text{ind}(k^*) > N-2$ case is that $j_k \geq j_{k^*} = 0, \forall k \neq k^*$. We can similarly obtain

$$\tilde{R}_T \leq 2K\Delta_{\max} + \sum_{k \neq k^*}^{K} \left(3c\Delta_k + 16(K-N+3)T_k(T)\sqrt{\frac{\ln T}{\frac{c}{n}-3}}\right).$$

Merging the above three cases, the expected dynamic regret upper bound is

$$\tilde{R}_T \leq 2K\Delta_{\max} + (K-1)\bar{\Delta}c\left(3 + \left\lceil \frac{K_0-N+2}{N-3} \right\rceil\right) + 16(K-1)(K-N+3)T\sqrt{\frac{n\ln T}{c-3n}},$$

which is minimized when

$$c = 3n + \left(\frac{8(K-N+3)T\sqrt{n\ln T}}{\bar{\Delta}(3 + \lceil \frac{K_0-N+2}{N-3} \rceil)}\right)^{\frac{2}{3}}.$$

If c is exactly $3n + \lceil \big(\frac{8(K-N+3)T\sqrt{n\ln T}}{\bar{\Delta}(3+\lceil \frac{K_0-N+2}{N-3} \rceil)}\big)^{\frac{2}{3}} \rceil$, it can be verified that $\tilde{R}_T = O\big(K^2 T^{\frac{2}{3}} (n\ln T)^{\frac{1}{3}}\big)$. □

H. Discussion on the Knowledge of the Patience Vector $\boldsymbol{m}$

For the following reasons, we assume that the value of $\boldsymbol{m}$ is known in advance in this paper:

1. In practice, $\boldsymbol{m}$ can come as a result of negotiation between the algorithm and arms in advance. We take the crowdsourcing scenario as our example. The system (algorithm) can make a promise of m_k for worker (arm) k such that in any time period of length m_k, the worker will surely have at least one job assignments. After a suitable value of $\boldsymbol{m}$ is confirmed, the proposed algorithms in this paper (FC-SE and FC-Entry) are able to ensure that this promise is never violated.

2. It is sufficient for the known $\boldsymbol{m}$ to only be a valid element-wise lower bound for the true patience vector. For the clarity of our discussion, say $\boldsymbol{m}'$ is the underlying true arm patience vector and $\boldsymbol{m}$ is an element-wise lower bound for $\boldsymbol{m}'$, i.e., $m_k \leq m'_k$ for any $k \in [K]$. The knowledge of any valid $\boldsymbol{m}$, instead of the exact $\boldsymbol{m}'$, is sufficient for the construction of a feasible cycle for the set of arms. This is because, if we repeat a feasible cycle constructed given $\boldsymbol{m}$, for any arm k, it is never continuously ignored for a duration of m_k time steps, thus arm k is also never continuously ignored for $m'_k \geq m_k$ time steps and it never leaves the game.

3. If no a-priori knowledge of arm patience $\boldsymbol{m}$ is accessible, we can prove that the bandit learning problem is intractable. Unlike learning the mean reward $\boldsymbol{\mu}$, learning the arm patience $\boldsymbol{m}$ is impractical, since obtaining partial information of $\boldsymbol{m}$ is accompanied by the risk of losing the optimal arm. Besides, in this scenario, all arms are initially indistinguishable for any algorithm. Intuitively, for any algorithm, there exists instances $(\boldsymbol{\mu}, \boldsymbol{m})$ such that the optimal arm leaves very early. In fact, we can formally prove that, without the knowledge of $\boldsymbol{m}$, any algorithm can incur unacceptable regret that is linear in the time horizon T:

Proposition H.1. *Given any algorithm A that only takes the historical observations $(a_1, r_1, ..., a_{t-1}, r_{t-1})$ as input at the beginning of time slot t and outputs an action a_t. The algorithm observes reward r_t at time t. Then there exists a family of problem instances such that the expected regret R_T satisfies $R_T = \Omega(T)$.*

Proof of Proposition H.1. Note that the algorithm A has no access to the arm patience m. Assume $K > 2$. There exists $i' \in [K]$ such that $\Pr(a_1 = i') \geq 1/K$, since otherwise $1 = \sum_{i=1}^{K} \Pr(a_1 = i) < \sum_{i=1}^{K} 1/K = 1$. Similarly, there exists $j' \in [K]$ such that $\Pr(a_2 = j' | a_1 = i', r_1 = 0) \geq 1/K$. Note that i' can be equal to j'. We construct a problem instance as follows: Select $k' \in [K]$ that satisfies $k' \neq i', k' \neq j'$. Set $\mu_{k'} = 1, m_{k'} = 2$. For $k \neq k'$, set $\mu_k = 0, m_k = T$. Set the reward noise to be 0 almost surely. That is, $r_t = \mu_{a_t}$ almost surely.

Obviously, arm k' is an impatient optimal arm. If the algorithm pulls $a_1 \neq k'$ and $a_2 \neq k'$, the optimal arm leaves at the end of time slot $t = 2$. We have

$$R_T \geq \Pr(a_1 \neq k', a_2 \neq k')T \geq \Pr(a_1 = i', a_2 = j')T.$$

We observe that

$$\begin{aligned}\Pr(a_2 = j' | a_1 = i') =& \sum_r \Pr(a_2 = j' | a_1 = i', r_1 = r) \Pr(r_1 = r | a_1 = i') \\ =& \Pr(a_2 = j' | a_1 = i', r_1 = 0)\end{aligned}$$

since $\mu_{i'} = 0$. Now we see

$$\begin{aligned}\Pr(a_1 = i', a_2 = j') =& \Pr(a_1 = i') \Pr(a_2 = j' | a_1 = i') \\ =& \Pr(a_1 = i') \Pr(a_2 = j' | a_1 = i', r_1 = 0) \\ \geq& 1/K^2.\end{aligned}$$

As a result, we have shown that $R_T \geq T/K^2$. □