Dare not to Ask: Problem-Dependent Guarantees for Budgeted Bandits

Nadav Merlis
Technion – Institute of Technology
merlis@campus.technion.ac.il

Yonathan Efroni
Microsoft Research, New York
yefroni@microsoft.com

Shie Mannor
Technion – Institute of Technology
Nvidia Research, Israel
shie@ee.technion.ac.il

Abstract

We consider a stochastic multi-armed bandit setting where feedback is limited by a (possibly time-dependent) budget, and reward must be actively inquired for it to be observed. Previous works on this setting assumed a strict feedback budget and focused on not violating this constraint while providing problem-independent regret guarantees. In this work, we provide problem-dependent guarantees on both the regret and the asked feedback. In particular, we derive problem-dependent lower bounds on the required feedback and show that there is a fundamental difference between problems with a unique and multiple optimal arms. Furthermore, we present a new algorithm called BuFALU for which we derive problem-dependent regret and cumulative feedback bounds. Notably, we show that BuFALU naturally adapts to the number of optimal arms.

1 Introduction

In the stochastic multi-armed bandit (MAB) problem [Robbins, 1952], an agent repeatedly selects actions (‘arms’) from a finite set and obtains their rewards, generated independently from arm-dependent distributions. The goal is to maximize the cumulative obtained reward, or, alternatively, minimize the regret. This setting has been extensively studied and its extensions are ubiquitous. In particular, many of its variants suggest different feedback models – structured [Chen et al., 2016a], partial, delayed and/or aggregated [Pike-Burke et al., 2018]. Then, once the feedback model is fixed, rewards are always observed through this model.

Nevertheless, such models ignore a key element in sequential decision-making: whatever the feedback model is, asking to observe rewards usually comes at some cost. In some cases, the cost is evident from the setting, e.g., if experimentation is required or the reward is manually labeled by an expert. In other settings, the feedback cost is more subtle. For example, in recommendation systems, incessantly asking to rate recommendations will aggravate many users. Therefore, it is natural to allow agents to decide whether they ask for feedback and design algorithms to ask for feedback with care.

One notable work that took such an approach is [Efroni et al., 2021]. There, agents must ask to observe rewards and cannot query more times than a given – possibly time-dependent and adversarial – reward-querying budget. Such a budget constraint allows to dynamically control the tradeoff between performance and querying ramifications. Then, the authors presented the Confidence-Budget Matching (CBM) principle and used it to adapt existing bandit and reinforcement learning (RL) algorithms to this budgeted setting. For the resulting agents, they proved tight problem-independent regret bounds as a function of the given budget.

Yet, this work has two notable shortcomings. First, its guarantees are problem-independent, namely, depend only on the problem size and given budget. This is in contrast to problem-dependent guarantees that depend on the specific instance faced by the agent and are generally more desirable. Second, and more importantly, the budget is treated as a hard constraint, and as long as this constraint is not violated, there is
no effort to refrain from asking for unnecessary feedback. Instead, agents should query rewards only when it is beneficial, regardless of the total budget. This is especially evident when relying on human feedback, as we now illustrate.

**Example 1** (restaurant recommendation problem). Consider a restaurant recommendation problem that learns from user-rankings (‘feedback’). Repeatedly asking for rankings will annoy users, so it is natural to cap the ranking requests (‘budget’) with an initially low cap that gradually increases to allow learning. However, expanding the cap whenever possible is ill-advised, as every query irritates the user, so asking for too many queries will do more harm than good. This can be modeled as a ‘soft’ cost for feedback queries that we never observe and cannot be directly translated to rewards, but greatly impacts the user experience. Then, we should avoid querying when not necessary, even if the hard budget allows it. Oftentimes, we will be willing to pay a small price in the recommendations quality (‘higher regret’) if it allows us to drastically reduce the number of queries, to avoid the unknown querying soft cost.

In this work, we address both shortcomings of Efroni et al. [2021] and analyze problem-dependent budget-conserving bounds for stochastic MAB problems. We first characterize problem-dependent lower bounds on the number of observations required for algorithms to perform well. Remarkably, we show a clear separation — which is absent in the usual bandit setting — between problems where the optimal (highest rewarding) arm is unique and problems with multiple optimal arms. In the former, we show that it suffices to query the optimal arm a logarithmic number of times, while in the latter, we show that any logarithmic rate can never suffice for well-behaved algorithms and problems. Then, we present the BuFALU algorithm and bound both its regret and expected number of queries using problem-dependent quantities. Notably, we show that BuFALU naturally identifies problems with a unique optimal arm and avoids wasting its querying budget unnecessarily. Finally, we numerically illustrate that BuFALU adapts to the number of optimal arms as the theoretical bounds suggest, while all reasonable baselines fail to do so.

### 1.1 Related Work

Surprisingly, sequential learning with feedback constraints has been, to a large extent, unexplored. To the best of our knowledge, this work is the first to study problem-dependent anytime guarantees for the MAB problem in the presence of time-dependent feedback constraints. Closely related to our work is the sequential budgeted learning framework in Efroni et al. [2021]. There, a non-decreasing cumulative budget is revealed to an agent at the beginning of each round, with the goal of minimizing the regret while satisfying the budget constraint almost surely. However, Efroni et al. [2021] focus on problem-independent regret guarantees, as opposed to problem-dependent guarantees as we consider. Furthermore, we alleviate their almost-sure restriction on the budget consumption by providing guarantees on the expected number of queries.

Another related setting is the problem of MAB with additional observations Yun et al. [2018]. There, agents are allowed to ask for observations from arms that were not played, but such queries are limited by a time-dependent budget. Importantly, the rewards of played arms are always observed. In contrast, we limit the number of observations from played arms. Thus, the settings are complementary. Finally, previous works studied adversarial settings where observing rewards incurs a cost Seldin et al. [2014], or where the interaction ends when the budget is exhausted Badanidiyuru et al. [2013]. These models substantially differ from ours, as explained in Efroni et al. [2021].

### 2 Setting

Our setting is inspired by the sequential budgeted learning model of Efroni et al. [2021]. In our problem, an agent (bandit strategy) faces $K$ arms (actions), each characterized by a reward distribution $\nu_a$ of expectation $\mu_a$. We refer to $\mathcal{L} = \{\nu_a\}_{a \in [K]}$ as the bandit instance, where $[K] \triangleq \{1, \ldots, K\}$, and denote the set of all valid reward distributions by $\mathcal{D}$ (e.g., all distributions supported by $[0, 1]$). When not clear from the context, we denote the expectation w.r.t. a specific bandit instance by $\mathbb{E}_\mathcal{L}$. The optimal reward of a bandit instance is $\mu^* = \max_{a \in [K]} \mu_a$, and we define the set of all optimal arms as $\mathcal{A}_* = \{a : \mu_a = \mu^*\}$. The suboptimality gap of an arm $a$ is $\Delta_a = \mu^* - \mu_a$, and we denote the maximal and minimal gaps by $\Delta_{\text{max}} = \max_a \Delta_a$ and $\Delta_{\text{min}} = \min_{a : \Delta_a > 0} \Delta_a$, respectively.
At each round \( t \geq 1 \), the agent plays a single arm \( a_t \in [K] \). Then, the arm generates a reward \( R_t \sim \nu_{a_t} \), independently at random of other rounds. However, to observe this reward, the agent must actively query it by setting \( q_t = 1 \); otherwise, the reward is not observed and \( q_t = 0 \). For brevity, we say that the agent always observes \( R_t \). We denote the number of times an arm was played up to round \( t \) by \( n_t(a) = \sum_{s=1}^t \mathbb{1}\{a_s = a\} \), where \( \mathbb{1}\{\cdot\} \) is the indicator function, and the number of times it was queried by \( n^q_t(a) = \sum_{s=1}^t \mathbb{1}\{a_s = a, q_s = 1\} \). Notice that for an arm to be queried, it must first be played, and thus \( n^q_t(a) \leq n_t(a) \). We similarly denote the total number of queries up to time \( t \) by \( B^q(t) = \sum_{a=1}^K n^q_t(a) \) and sometimes limit it by some querying budget \( B(t) \). We assume that querying an arm incurs a unit querying-cost but remark that all bounds can be easily extended to arm-dependent costs. Finally, we define the empirical mean of an arm \( a \), based on observed samples up to round \( t \), by \( \hat{\mu}_t(a) = \frac{1}{n^q_t(a)} \sum_{s=1}^t R_s \mathbb{1}\{a_s = a, q_s = 1\} \) and say that \( \hat{\mu}_t(a) = 0 \) if \( n^q_t(a) = 0 \).

In this work, we evaluate agents by their regret, namely, the expected difference between the optimal reward and the reward of the played arms:

\[
\text{Reg}(T) = \mathbb{E}\left[ \sum_{t=1}^T (\mu^* - \mu_{a_t}) \right] = \mathbb{E}\left[ \sum_{t=1}^T \Delta_{a_t} \right]
\]

### 3 Lower Bounds

In this section, we present lower bounds on the number of queries we must take from arms for the agent to ‘behave well’ (for different notions of good behavior). Importantly, we will see a distinctly different behavior of the lower bounds when there is a unique or multiple optimal arms. This will encourage us to design an algorithm that adapts to both scenarios, as we do in the next section.

We require a few additional notations. Let \( \mathcal{A}_t(\nu) \) be the set of optimal arms in instance \( \nu \). Also, denote the Kullback-Leibler (KL) between any two distributions \( \nu_{a} \) and \( \nu'_{a} \) by \( \text{KL}(\nu_{a}, \nu'_{a}) \) and the KL divergence between two Bernoulli random variables of expectations \( p, q \) by

\[
\text{kl}(p, q) = p \ln \frac{p}{q} + (1 - p) \ln \frac{1 - p}{1 - q}.
\]

Then, we define

\[
\mathcal{K}_{\inf}^+(\nu, \mu, \mathcal{D}) = \inf\{\text{KL}(\nu, \nu') : \nu' \in \mathcal{D}, \mathbb{E}[\nu] > \mu\},
\]

\[
\mathcal{K}_{\inf}^-(\nu, \mu, \mathcal{D}) = \inf\{\text{KL}(\nu, \nu') : \nu' \in \mathcal{D}, \mathbb{E}[\nu] < \mu\},
\]

where the infimum over an empty set is \(+\infty\). If the infimum is zero, we let its inverse be \(+\infty\). Intuitively, \( \mathcal{K}_{\inf}^+(\nu_{a}, \mu, \mathcal{D}) \) represents the distance between a distribution \( \nu_{a} \) to the closest distribution in \( \mathcal{D} \) of expectation higher than \( \mu \). Then, to distinguish between \( \nu_{a} \) and any distribution of expectation larger than \( \mu \), we would require a number of samples that is inversely proportional to \( \mathcal{K}_{\inf}^+(\nu_{a}, \mu, \mathcal{D}) \). Similarly, \( \mathcal{K}_{\inf}^- \) can be related to the closest distribution of a lower expectation.

Finally, let \( \{U_t\}_{t \geq 0} \) be a sequence of i.i.d. uniform random variables that encompass the internal randomization of agents. Then, a bandit strategy maps the history \( I_t = (U_0, a_1, q_1, R_1 \cdot q_1, U_1, \ldots, a_t, q_t, R_t \cdot q_t, U_t) \) into a next action \( a_{t+1} \) and querying rule \( q_{t+1} \). Under this notation, Efroni et al. [2021] proved the following result:

**Lemma 1.** For any \( T \geq 1 \), any \( \sigma(I_T) \)-measurable random variable \( Z \) with values in \([0, 1]\) and any two bandit instances \( \nu \) and \( \nu' \), it holds that

\[
\sum_{a=1}^K \mathbb{E}[n^q_T(a)]\text{KL}(\nu_{a}, \nu'_{a}) \geq \text{kl}(\mathbb{E}[Z], \mathbb{E}[Z])
\]

As was shown in Garivier et al. [2019], a similar inequality allows elegantly deriving lower bounds under various regularity assumptions on the bandit strategy.
### 3.1 Asymptotic Lower Bounds

Probably the most common assumption for a bandit strategy is that for any bandit instance, the regret of the strategy is asymptotically sub-polynomial. We call such strategies consistent:

**Definition 1.** A bandit strategy is called consistent w.r.t. $D$ if for any instance $\nu \in D$, any suboptimal arm $a \notin A_*(\nu)$ and any $\alpha \in (0, 1)$ it holds that $n_T(a) = o(T^\alpha)$.

For consistent strategies, the following holds:

**Theorem 1.** Let $n_\infty^a(a) = \liminf_{T \to \infty} \frac{E[n_T^a(a)]}{\ln T}$ for $a \in [K]$. If the bandit strategy is consistent w.r.t. $D$, then for any bandit instance $\nu \in D$, the following hold:

1. For any suboptimal arm $a \notin A_*(\nu)$,
   \[
   n_\infty^a(a) \geq \frac{1}{\mathcal{K}_\text{inf}(v_a, \mu_*, D)}. \tag{2}
   \]

2. Assume that $a^*$ is the unique optimal arm and let $a \neq a^*$ be a suboptimal arm. Also, denote the maximal suboptimal reward by $\mu^* = \max_{a \neq a^*} \mu_a$. Then $n_\infty^a(a^*) \geq \frac{1}{\mathcal{K}_\text{inf}(v_{a^*}, \mu_*, D)}$ and for any $\mu \in [\mu^*, \mu^*]$,
   \[
   n_\infty^a(a) + n_\infty^a(a^*) \geq \frac{1}{\max\{\mathcal{K}_\text{inf}(v_\mu, \mu, D), \mathcal{K}_\text{inf}(v_{a^*}, \mu, D)\}}. \tag{3}
   \]

3. Assume that there are at least two optimal arms. Moreover, assume that (1) for all optimal arms $a \in A_*(\nu)$, $\mathcal{K}_\text{inf}(v_a, \mu^*, D) = 0$, or, alternatively, (2) there are at least two optimal arms for which $\mathcal{K}_\text{inf}(v_\mu, \mu^*, D) = 0$. Then $n_\infty^a(a) = \infty$ for some optimal arm $a \in A_*(\nu)$.

The proof is in Appendix A. Notice that the bounds of the theorem are asymptotic, namely, for large enough $T$, an action $a$ is roughly queried $n_\infty^a(a) \cdot \ln T$ times. Importantly, recall that $n_T^a(a) \leq n_t(a)$, so all results also hold for the number of plays. The theorem is divided into three parts. The first part is a natural extension of the classical problem-dependent lower bound for MABs [Lai and Robbins, 1985, Burnetas and Katehakis, 1996] and emphasizes that it is not enough to sufficiently play suboptimal arms, but we rather must sufficiently query them. The second and third parts discuss the querying requirements from optimal arms when there is a unique or multiple optimal arms, respectively. This comes in stark contrast to the classical lower bounds that disregard querying, as playing optimal arms does not incur regret and can thus be ignored.

When there is a unique optimal arm $a^*$, the result first states that it must be distinguished from the highest suboptimal arm $a$. Yet, Equation (3) implies that by itself, this does not suffice. Instead, for any suboptimal arm $a$, both $a$ and $a^*$ must be sufficiently queried to separate them; namely, identifying that $a^*$ is better than $a$. To see this, consider the (typical case) where $\mathcal{K}_\text{inf}(\nu_\mu, D)$ and $\mathcal{K}_\text{inf}(\nu, D)$ are continuous in $\mu$ and equal zero if $\mathbb{E}[\nu] = \mu$. Also, notice that $\mathcal{K}_\text{inf}(\mathcal{K}_\text{inf})$ decreases (increases) in $\mu$. Then, there exists $\mu_0 \in (\mu_*, \mu^*)$ such that $\mathcal{K}_\text{inf}(\nu_{\mu_0}, D) = \mathcal{K}_\text{inf}(\nu_{a^*}, \mu_0, D)$, and this choice maximizes the r.h.s. of (3). In this case, a reasonable way to match the lower bound is to ensure that both $n_\infty^a(a)$ and $n_\infty^a(a^*)$ are roughly equal to $(\mathcal{K}_\text{inf}(\nu_{\mu_0}, D))^{-1}$. This separates the optimal arm $a^*$ from distributions $\{\nu \in D : \mathbb{E}[\nu] < \mu_0\}$ and the suboptimal arm $a$ from $\{\nu \in D : \mathbb{E}[\nu] > \mu_0\}$. Concretely, if $D$ is the set of all Gaussian distributions of unit variance, then $\mu_0 = \mu_a + \Delta_a/2 = \mu^* - \Delta_a/2$, and we should estimate both $a$ and $a^*$ up to a precision of $\Delta_a/2$.

Finally, the last part of Theorem 1 treats problems with multiple optimal arms. Specifically, it states that a logarithmic budget might suffice only if $\mathcal{K}_\text{inf}(\nu_\mu, \mu^*, D) = 0$ for at most a single optimal arm and $\mathcal{K}_\text{inf}(\nu_\mu, \mu^*, D) > 0$ for at least one optimal arm. Notably, for standard distribution sets $D$, for any value of $\mu^*$, either $\mathcal{K}_\text{inf}(\nu_\mu, \mu^*, D) = 0$ or $\mathcal{K}_\text{inf}(\nu_\mu, \mu^*, D) = 0$ for all $\nu \in D$ with $\mathbb{E}[\nu] = \mu^*$, i.e., at least one of the conditions hold and multiple optimal arms should be queried super-logarithmically. Intuitively, when either of the conditions hold, it is impossible to determine if arms are optimal or near-optimal with arbitrarily small gaps by logarithmic querying. We illustrate it in a concrete example of a distribution set $D$ in Appendix A.2.
To summarize, Theorem 1 creates a remarkable distinction between budget-constrained problems with a unique optimal arm, where a logarithmic budget is adequate, and problems with multiple optimal arms, where consistent algorithms must query super-logarithmically. Similar phenomena do not exist in classical MAB problems and a key contribution of this theorem is the refined characterization of conditions for it to occur.

### 3.2 Querying Costs and Budget Conservation

We now discuss the implications of the asymptotic lower bounds when querying rewards incur a cost \( c > 0 \). The cost might be known (e.g., payment for labeling) or unknown (e.g., user irritation in recommender systems). In this setting, it is natural to modify the regret to be

\[
\text{Reg}^q(T) = \mathbb{E} \left[ \sum_{t=1}^{T} (\mu^* - \mu_{a_t}) \right] + \mathbb{E} \left[ \sum_{t=1}^{T} c 1\{q_t = 1\} \right]
\]

\[
= \sum_{a=1}^{K} (\Delta_a \mathbb{E}[n_T(a)] + c \mathbb{E}[n^q_T(a)])
\]

\[
\geq \sum_{a=1}^{K} (\Delta_a + c) \mathbb{E}[n^q_T(a)],
\]

where the inequality is since actions must be played to be queried (\( n^q_T(a) \leq n_T(a) \)). Assume that the regret is sub-polynomial for any instance in \( D \); in particular, the strategy is consistent and Theorem 1 holds.

Now, if the optimal arm is unique, then strategies must separate the optimal arm from any suboptimal arm (by the second part of Theorem 1). However, doing so with super-logarithmic queries leads to suboptimal regret due to the querying costs. Thus, optimal algorithms should query all arms logarithmically, including the optimal arm. The best balance between queries from the optimal arm and increased plays from suboptimal arms depends on the values of \( c \) and the gaps. However, reward degradation is unavoidable for any \( c > 0 \).

On the other hand, if there are multiple optimal arms and the conditions of the last part of Theorem 1 hold, then optimal arms must be queried super-logarithmically for a strategy to be consistent. Therefore, no strategy can achieve logarithmic regret, and every strategy must either suffer from high querying costs or low rewards.

### 3.3 Lower Bounds for Small Budget

Finally, we discuss the best possible performance in the limit of small budgets. To derive these bounds, we require the bandit strategy to be better-than-uniform:

**Definition 2.** A bandit strategy is called better-than-uniform on \( D \) if for any instance \( \nu \in D \) and any \( T \geq 1 \),

\[
\sum_{a \in A^*_\nu} \mathbb{E}_\nu[n_T(a)] \geq \frac{|A^*_\nu|}{K} T .
\]

This definition slightly differs from the one in Garivier et al. [2019], which requires all optimal arms to be played in expectation at least \( T/K \) times. This does not affect the result but allows focusing on playing a specific optimal arm, which better fits budgeted settings. For such strategies, the following holds (see proof in Appendix A.3, which resembles the one of Theorem 2 in Garivier et al. [2019]):

**Proposition 2.** For any bandit instance \( \nu \in D \), any strategy that is better than uniform on \( D \), any arm \( a \notin A^*_\nu \) and any \( T \geq 1 \), it holds that

\[
\mathbb{E}_\nu[n_T(a)] \geq \frac{T}{K} \left( 1 - \sqrt{2 \min \{K \mathbb{E}_\nu[n^q_T(a)], T\} K^+_{int}(\nu_a, \mu^*, D)} \right).
\]
If $K^+_{\inf}(\nu_a, \mu^*, D) = \infty$, we define the r.h.s. to be zero. Specifically, if $K^+_{\inf}(\nu_a, \mu^*, D) < \infty$ and $\mathbb{E}_{\nu}[n_T^2(a)] \leq \frac{1}{8K^+_{\inf}(\nu_a, \mu^*, D)}$, then $\mathbb{E}_{\nu}[n_T(a)] \geq T/(2K)$.

This proposition has an important implication on the regret when feedback is limited by a budget:

**Corollary 3.** Assume that for any $t \geq 1$ and $a \in [K]$, $\mathbb{E}_{\nu}[n_T^2(a)] \leq B_a(t)$ for some positive nondecreasing budget $\{B_a(t)\}_{t \geq 1}$. Also let $B_a^{-1}(x) = \sup\{t \in \mathbb{N} : B_a(t) \leq x\}$ for $x \geq B(1)$ and otherwise $B_a^{-1}(x) = 0$. Then, for any instance $\nu \in D$ and any better-than-uniform bandit strategy on $D$,

$$\forall T \geq \max_{a \notin A^*(\nu)} B_a^{-1}\left(\frac{1}{8K^+_{\inf}(\nu_a, \mu^*, D)}\right), \quad \text{Reg}(T) \geq \sum_{a \notin A^*(\nu)} \frac{\Delta_a}{2K} B_a^{-1}\left(\frac{1}{8K^+_{\inf}(\nu_a, \mu^*, D)}\right).$$

The proof is in Appendix A.3. We remark that although Corollary 3 is mainly useful with budget constraints, one can apply it with any instance $\nu \in D$ and any better-than-uniform bandit strategy on $D$.

**Example 2 (Consequences of budget profiles).** Let the set of all possible arm distributions $D$ be the set of Gaussian distributions with unit variance. In this case, it is easily verified that for all $a \in [K]$, $K^+_{\inf}(\nu_a, \mu^*, D) = \frac{(\mu^* - \mu_a)^2}{2} = \frac{\Delta_a^2}{2}$. Then, by Corollary 3 for any $T \geq \max_{a \notin A^*(\nu)} B_a^{-1}\left(\frac{1}{4K\Delta_a^2}\right)$, the regret is lower bounded by

$$\text{Reg}(T) \geq \sum_{a \notin A^*(\nu)} \frac{\Delta_a}{2K} B_a^{-1}\left(\frac{1}{4K\Delta_a^2}\right).$$

Consider the following budget limitation functional:

1. If the per-arm budget is polynomial $B_a(T) = T^\alpha/K$ for $\alpha \in [0, 1]$, then the lower bound is $\text{Reg}(T) = \Omega\left(\sum_{a \notin A^*(\nu)} \Delta_a^{1-\frac{\alpha}{2}}\right)$, i.e., the lower bound is inversely polynomial in the gaps.

2. If the per-arm budget is poly-log and gap-oblivious, i.e., $B_a(T) = (\ln T)^\alpha/K$, then the lower bound is exponential in the inverse-gaps: $\text{Reg}(T) = \Omega\left(\sum_{a \notin A^*(\nu)} \Delta_a \exp\left\{\Delta_a^{-2/\alpha}\right\}\right)$.

3. If the per-arm budget is logarithmic but gap-aware, i.e., $B_a(T) = (\ln T)/\Delta_a^2$, then the lower bound is linear in the gaps: $\text{Reg}(T) = \Omega\left(\sum_{a \notin A^*(\nu)} \Delta_a/K\right)$. Typically, $\Delta_a \leq 1$, and the bound is effectively constant. Notably, it suffices to know a lower bound on the gaps: if $\Delta_a \geq \epsilon$ for all $a \notin A^*(\nu)$, then a total budget of $B(T) = (K \ln T)/\epsilon^2$ is adequate.

Typically, agents do not know the value of the gaps. Thus, although a poly-logarithmic budget might be adequate asymptotically, it leads to a prohibitively large regret – exponential in the gaps. Notably, this is the case even when there exists a unique optimal arm.

## 4 Upper Bounds

In the previous section, we showed that algorithms should allocate enough (possibly super-logarithmic) querying budget to optimal arms. We also showed that the regret is linear until we sufficiently query suboptimal arms. Finally, in the important case of a unique optimal arm, we showed that the optimal arm should be adequately separated from all suboptimal arms but argued that it should be sampled logarithmically to avoid querying costs. In this section, we leverage these insights to design an algorithm that conservatively queries rewards. For simplicity, we focus on problems with rewards bounded in $[0, 1]$, but in Appendix B, we prove our results for any confidence intervals that follow some mild assumptions (including Bernstein bounds).

To be more feedback-aware, we adopt a confidence-based approach. Namely, we assume that each arm $a$ is equipped with a confidence interval of width $CI_t(a) = UCB_t(a) - LCB_t(a) \geq 0$, such that the true mean of arms is within the confidence interval with a sufficiently high probability, i.e., $\mu_a \in [LCB_t(a), UCB_t(a)]$.?
Algorithm 1 Budget-Feedback Aware Lower Upper Confidence Bound (BuFALU)

1: Define: $UCB_t(a) = \hat{\mu}_{t-1}(a) + \sqrt{\frac{3\ln t}{2n_{a,t-1}(a)}}$, $LCB_t(a) = \hat{\mu}_{t-1}(a) - \sqrt{\frac{3\ln t}{2n_{a,t-1}(a)}}$ ('Hoeffding')
2: for $t = 1, ..., K$ do
3:   Play $a_t = t$ and ask for feedback ($q_t = 1$)
4:   Observe $R_t$ and update $n_t^a(a_t), \hat{\mu}_t(a_t)$
5: end for
6: for $t = K + 1, ..., T$ do
7:   Observe $\epsilon(t) \geq 0$
8:   Set $l_t \in \arg\max_a LCB_t(a), u_t \in \arg\max_a UCB_t(a)$ and $c_t \in \arg\max_{a \in \{l_t, u_t\}} CI_t(a)$
9:   if $UCB_t(u_t) \leq LCB_t(l_t)$ or $UCB_t(c_t) - LCB_t(l_t) \leq \epsilon(t)$ then
10:      Play $a_t = l_t$; do not ask for feedback ($q_t = 0$)
11:   else
12:      Play $a_t = c_t$; ask for feedback ($q_t = 1$)
13:      Observe $R_t$ and update $n_t^a(a_t), \hat{\mu}_t(a_t)$
14: end if
15: end for

Importantly, confidence intervals can be used to quantify the suboptimality of actions. For example, for any arm $a$, we can upper bound its suboptimality gap w.h.p. by

$$\Delta_a = \mu^* - \mu_a \leq UCB_t(a^*) - LCB_t(a) \leq \max_{a'} UCB_t(a') - LCB_t(a),$$

where $a^*$ is some optimal arm. Specifically, if $a \in \arg\max_a UCB_t(a')$, then $\Delta_a \leq CI_t(a)$. In general, by controlling the widths of confidence intervals, we can identify good actions and bound their suboptimality. However, narrowing these intervals requires reward samples, which are limited at our setting.

One approach to face the budget limitation is to try shrinking confidence intervals up to a fixed certainty level, determined by the querying budget. This can be done using the confidence-budget matching mechanism [CBM, Efroni et al., 2021], which plays $a_t \in \arg\max_a UCB_t(a)$ and queries for feedback only if $CI_t(a_t) \geq \sqrt{K/B(t)}$ for a querying budget $B(t)$. This leads to favorable problem-independent regret without violating the budget constraint. Sadly, CBM miserably fails to conserve budget when the optimal arm is unique. This can be easily observed in the limit of infinite budget ($B(t) \to \infty$), where the algorithm always queries, even when the optimal arm is unique.

To solve this issue, we refine the CBM confidence-controlling scheme and improve the action-selecting mechanism. For action selection, inspired by the lower bounds, we aim to separate the confidence interval of a unique optimal arm $a^*$ from the rest of the arms, i.e., $LCB_t(a^*) \geq UCB_t(a)$ for all suboptimal $a$. If we succeeded in doing so, we can safely declare that $a^* = \arg\max_a LCB_t(a) \triangleq l_t$ and play $l_t$ without reward querying. Letting $u_t \in \arg\max_{a \neq l_t} UCB_t(a)$, this case happens if $UCB_t(u_t) \leq LCB_t(l_t)$. When this condition does not hold, we would like to separate $u_t$ from $l_t$ by actively shrinking the confidence intervals of both arms. An efficient way to do so is to play (and query) the arm with the wider confidence interval, namely,

$$c_t \in \arg\max_{a \in \{u_t, l_t\}} CI_t(a) \quad \text{for} \quad l_t \in \arg\max_a LCB_t(a) \quad \text{and} \quad u_t \in \arg\max_{a \neq l_t} UCB_t(a). \quad (4)$$

However, this separation might be very costly in queries and is impossible if there are multiple optimal arms. Thus, as in CBM, we utilize the confidence intervals to regulate the number of queries. We suggest a more abstract scheme than CBM and query for reward only if $UCB_t(c_t) - LCB_t(l_t) > \epsilon(t)$ for some (possibly adversarial) $\epsilon(t) \geq 0$. Note that $UCB_t(c_t) - LCB_t(l_t) \leq CI_t(l_t)$, so this condition refines the CBM querying rule and requires $c_t$ to be sufficiently better than $l_t$. Particularly, letting $\epsilon(t) \approx \sqrt{K/B(t)}$ results in a querying condition similar to CBM. Yet, using $\epsilon(t)$ instead of $B(t)$ provides additional flexibility to the framework; for example, setting $\epsilon(t) = \infty$ blocks querying from certain rounds and $\epsilon(t) = 0$ allows free queries.

Lastly, if $UCB_t(c_t) - LCB_t(l_t) \leq \epsilon(t)$, we revert to playing $l_t$ without reward querying. We later show that doing so is ‘safe’, in a sense that $\Delta_l \leq \epsilon(t)$ (Lemma 3 in Appendix B.1). Combining both playing and
 querying schemes leads to the Budget-Feedback Aware Lower-Upper Confidence Bound algorithm (BuFALU), depicted in Algorithm 1. Before stating the regret and querying guarantees of BuFALU, we present two notable quantities on which the bounds will depend:

\[ L_c(T, \Delta) = \sum_{t=1}^{T} \mathbb{1}\{c(t) \geq \Delta\}, \quad \bar{N}(T, \Delta) = \max_{t \in [T]} \frac{6 \ln t}{\max \{\Delta^2, c^2(t)\}}. \]

The first quantity \( L_c(T, \Delta) \) counts the number of rounds that \( c(t) \) exceeds a fixed confidence level \( \Delta \) until time \( T \). To understand it, keep in mind that when a reward is not queried, we play \( a_t = l_t \) with the guarantee that \( \Delta_{l_t} \leq c(t) \) (by Lemma 5). In turn, at such rounds, we might play any arm \( a \) for which \( \Delta_a \leq c(t) \). Thus, \( L_c(T, \Delta_a) \) represents the maximal number of rounds that a suboptimal arm \( a \) can be played when reward are not queried. A notable case is when \( c(t) \) is nonincreasing, and then \( L_c(T, \Delta) \) can be conveniently bounded by \( L_c(T, \Delta) \leq \epsilon^{-1}(\Delta) \equiv \sup \{t \geq 1 : c(t) \geq \Delta\} \).

To understand \( \bar{N} \), recall that \( n_{l-1}^{\Delta}(a) = \frac{6 \ln t}{\mu^2} \) is the number of samples required for the confidence interval \( CI_t(a) \) to be smaller than \( \mu \). Thus, \( \bar{N}(T, \Delta) \) represents the maximal number of samples required to shrink the confidence intervals to a confidence level of \( \max \{c(t), \Delta\} \). In particular, for any suboptimal arm \( a \) and \( \Delta = \Delta_a \), this number of samples suffices to either identify that \( a \) is suboptimal (confidence smaller than \( \Delta_a \)) or to stop sampling it due to the confidence constraint. Importantly, if \( c(t) = \sqrt{6K \ln t/B(t)} \) for a positive nondecreasing budget \( B(t) \) (or, alternatively, if \( c(t) \) is nonincreasing), see that \( \bar{N}(T, \Delta) = \frac{6 \ln T}{\max \{\Delta^2, c^2(t)\}} \).

Then, we get that \( \bar{N}(T, 0) \leq B(T)/K \), so this term ensures that the budget constraint is never violated. We now state the performance bounds for BuFALU.

**Theorem 2.** Assume that the rewards are bounded in \([0, 1]\). Also, let \( T \geq 1 \) and assume that \( \{c(t)\}_{t \in [T]} \) is some nonnegative sequence. Then, when running Algorithm 1 the following hold

1. For all \( a \in [K] \), it holds that \( n_{l-1}^{\Delta}(a) \leq \bar{N}(T, 0) + 1 \).

2. If there are multiple optimal arms \( |\mathcal{A}_*| > 1 \), then

\[
\text{Reg}(T) \leq \sum_{a \notin \mathcal{A}_*} \Delta_a \left( \bar{N}(T, \Delta_a) + L_c(T, \Delta_a) \right) + 3K \Delta_{\max},
\]

\[
\mathbb{E}[B]^2(T) \leq \sum_{a \notin \mathcal{A}_*} \bar{N}(T, \Delta_a) + |\mathcal{A}_*| \bar{N}(T, 0) + 3K.
\]

3. If the optimal arm \( \mathcal{A}_* \) is unique, then

\[
\text{Reg}(T) \leq \sum_{a \neq \mathcal{A}_*} \Delta_a \left( \bar{N} \left( T, \frac{\Delta_a}{2} \right) + L_c(T, \Delta_a) \right) + 3K \Delta_{\min},
\]

\[
\mathbb{E}[B]^2(T) \leq \sum_{a \neq \mathcal{A}_*} \bar{N} \left( T, \frac{\Delta_a}{2} \right) + \bar{N} \left( T, \frac{\Delta_{\min}}{2} \right) + 3K.
\]

The proof can be found in Appendix B.1. The stated bounds hold for oblivious adversarially chosen \( c(t) \). If the adversary is adaptive, the same results hold by taking an expectation on all bounds.

The first part of the theorem provides strict querying guarantees and can be directly related to the budgeted setting in Efroni et al. 2021. Notably, for nondecreasing \( B(t) \) and \( c(t) = \sqrt{6K \ln t/B(t)} \), BuFALU asks up to \( B(T) + K \) queries. Yet, we emphasize that we allow any general choice of \( c(t) \geq 0 \), in sharp contrast to the budget-dependent thresholding in Efroni et al. 2021.

Next, we compare Theorem 2 to the lower bounds of Section 3 and for simplicity, assume that \( c(t) \) is nonincreasing. First, the per-arm query bound of \( B_a(t) = \tilde{O}(1/c^2(t)) \) implies that \( \epsilon^{-1}(\Delta_a) \approx B_a^{-1}(1/\Delta_a^2) \). Then, by bounding \( L_c(t, \Delta) \leq \epsilon^{-1}(\Delta) \), and for \( K_{\inf}(\nu, \mu^*, D) \approx \Delta_a^2 \), the regret term of \( \Delta_a L_c(T, \Delta_a) \) corresponds with the lower bound of Corollary 3. Moreover, for any \( c(t) \geq 0 \), the regret is logarithmic, and all suboptimal arms are logarithmically sampled. Specifically, when \( K_{\inf}(\nu, \mu^*, D) \approx \Delta_a^2 \), this matches the asymptotic lower bound up to absolute constants. However, if there are multiple optimal arms, they will be queried \( \tilde{O} \left( \frac{\ln T}{c^2(t)} \right) \) times each – super-logarithmically.
Finally, recall that UCB algorithms for the standard MAB problem typically leads to count (and query) bounds of $\frac{6 \ln T}{\Delta_a}$ (see, e.g., Theorem 2.1 of Bubeck et al. 2012 with $\alpha = 3$). Indeed, the suboptimal queries depend on $\tilde{N}(T, \Delta_a) \leq \frac{6 \ln T}{\Delta_a^2}$ when there are multiple optimal arms. In contrast, when the optimal arm is unique, the bounds depend on $\tilde{N}(T, \Delta_a/2) \approx 4 \tilde{N}(T, \Delta_a)$. This factor might be explained by the second part of the lower bound (Theorem 1): namely, to logarithmically query the unique optimal arm, it must be completely separated from all suboptimal arms. Then, the term $\Delta_a/2$ is the result of separating both the optimal arm and suboptimal arms from their middle point $(\mu_a + \mu^*)/2$.

We end this section by showing that BuFALU achieves the same problem-independent guarantees as CBM (see proof in Appendix B.2).

**Proposition 4.** When running Algorithm 1 with any sequence $\epsilon(t) \geq 0$, for any $T \geq 1$,

$$\text{Reg}(T) \leq 4\sqrt{6KT\ln T} + \sum_{t=1}^{T} \epsilon(t) + 3K\Delta_{\text{max}}.$$  

For $\epsilon(t) = \sqrt{\frac{6K\ln T}{B(t)}}$, we match the upper (and lower) bound of Efroni et al. 2021, namely $\text{Reg}(T) = O\left(\sqrt{KT\ln T} + \sum_{t=1}^{T} \sqrt{\frac{K\ln T}{B(T)}}\right)$. Moreover, our bound further extends this result to any sequence of $\epsilon(t) \geq 0$.

### 4.1 Numerical Illustration

![Figure 1: Evaluation of all algorithms on two deterministic instances. Left column: regret, right column: number of queries. In the top row, there is a unique optimal arm, while in the bottom one there are two. All algorithms were evaluated with $\epsilon(t) = t^{-1/4}$ on a single seed (deterministic problems)](image)

In this section, we present a simple numerical illustration of the behavior of BuFALU in the presence of a unique or multiple optimal arms. To best capture the difference between the scenarios, we evaluate the algorithm on two deterministic MAB instances. In the first instance, the optimal arm is unique (with
and there exists a single suboptimal arm (with $\mu_1 = 0$). The second instance is the same, except for an additional optimal arm. In the evaluation, we compare BuFALU to a few natural baselines. The first uses the same querying mechanism as BuFALU but sets $c_t \in \arg\max_a UCB_t(a)$. This baseline, called BuFAU, will allow us to understand the contribution of the action choice vs. the querying mechanism. The second baseline is CBM-UCB [Efroni et al., 2021], which has similar problem-independent guarantees as BuFALU. Finally, the third baseline is a greedy algorithm that receives a total budget as guaranteed by Theorem 2 (i.e., $\frac{K \ln T}{c^2(t)} + K$). If the budget was not exhausted, it plays (and queries) the maximal UCB; otherwise, it plays (without querying) the arm with the maximal empirical mean. All algorithms are evaluated with $\epsilon(t) = t^{-1/4}$ for 100,000 steps. Then, Theorem 2 guarantees that BuFALU cannot query any arm more than $\sim 20,000$ times. We refer to Appendix C.2 for a full description of the baselines and remark that similar evaluations in random 5-armed problems lead to the same insights (see Appendix C.2).

The simulation results are depicted in Figure 1. One immediate conclusion is that in these simple instances, all baseline algorithms behave roughly the same, where the only difference is that the greedy baseline completely exhausts its querying budget while all other baselines only exhaust it for optimal arms. Moreover, the simulated behavior of BuFALU validates the characterization of Theorem 2 when there are multiple optimal arms, BuFALU uses all available budget to query them and achieves the same regret and querying performance as the baselines. In contrast, when the optimal arm is unique, BuFALU only sparingly asks for feedback ($\sim \frac{1}{150}$, compared to the baselines), but its regret is larger by a factor of $4 -$ the same factor that we get by replacing $N(T, \Delta_k)$ with $N(T, \Delta_k/2)$. We emphasize that in the presence of soft querying costs, the query reduction is clearly worth this degradation. In fact, one can easily see that for any per-query cost of $c > 0.0025$ (namely, 0.25% of the optimal rewards), BuFALU outperforms all baselines.

Lastly, we see that the changes to the querying mechanism, applied in BuFAU, have a minor effect, so the key part is the choice of $c_t$, which allows separating a unique optimal arm from all other arms.

## 5 Summary and Discussion

In this work, we analyzed MAB problems where rewards must be queried to be observed. We proved lower bounds in this setting and highlighted the fundamental difference between problems with a unique and multiple optimal arms. We also presented BuFALU, to which we proved problem-dependent regret and querying bounds and showed that it naturally adapts to the number of optimal arms.

In the standard MAB setting, there are a few interesting directions for improving our results. First, although our analysis supports arbitrary sequences of $\epsilon(t)$, the maximization in $\bar{N}$ might be loose when $\epsilon(t)$ briefly drops. We believe that better characterizing these cases is imperative when working with general (possibly decreasing) budgets. Second, when the optimal arm is unique, our regret bounds degrade by a constant factor of 4. Yet, the lower bounds hint that this factor does not have to affect all arms (and might, for example, be present mainly in the term of the optimal arm). Improving this factor might require changes to the algorithm and/or confidence intervals and is left for future work. Finally, while we limited the number of queries from played arms, [Yun et al., 2018] allowed budget-limited observations from unplayed arms. Although combining the settings is natural, the individual solutions are very different, and we leave this for future work.

Moreover, we only tackled the standard MAB setting. Extending this work to other settings might lead to nontrivial challenges. First, in large or continuous problem, e.g., combinatorial bandits, [Chen et al., 2016a], linear bandits [Abbasi-Yadkori et al., 2011] and reinforcement learning [Jaksch et al., 2010, Azar et al., 2013, Simchowitz and Jamieson, 2019], the dichotomy to unique and multiple optimal arms might be too coarse. Then, it might be more relevant to characterize the behavior using the size or structure of the set of optimal actions. Specifically in RL, recent studies show that the presence of multiple optimal arms greatly affects the problem-dependent regret even with budget constraints [Xu et al., 2021, Tirinzoni et al., 2021], and it is worthwhile to further study it when feedback is limited. Second, adapting BuFALU to different domains is not always straightforward, and alternative approaches should be considered.

Finally, this work raises important questions in the low-budget regime (e.g., logarithmic budget). There, minimizing the regret seems hopeless in the presence of multiple optimal arms, and weaker optimality notions can be considered. One option is lenient regret criteria [Merlis and Mannor, 2020], which do not incur regret when playing near-optimal arms. Then, it might be possible to perform well even when $\epsilon(t)$ does not decrease
to zero, but this warrants further study.

Acknowledgments

This work was partially funded by the Israel Science Foundation under ISF grant number 2199/20. Nadav Merlis is partially supported by the Gutwirth Scholarship. YE is partially supported by the Viterbi scholarship, Technion.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216. IEEE, 2013.

Sébastien Bubeck, Nicole Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778, 2016a.

Yonathan Efroni, Nadav Merlis, Aadirupa Saha, and Shie Mannor. Confidence-budget matching for sequential budgeted learning. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376, 2011.

Aurélien Garivier, Hédi Hadiji, Pierre Menard, and Gilles Stoltz. Kl-ucb-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. *arXiv preprint arXiv:1805.05071*, 2018.

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

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

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

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Nadav Merlis and Shie Mannor. Lenient regret for multi-armed bandits. *arXiv preprint arXiv:2008.03959*, 2020.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113. PMLR, 2018.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.
Yevgeny Seldin, Peter Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *International Conference on Machine Learning*, pages 280–287. PMLR, 2014.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Andrea Tirinzoni, Matteo Pirotta, and Alessandro Lazaric. A fully problem-dependent regret lower bound for finite-horizon mdps. *arXiv preprint arXiv:2106.13013*, 2021.

Haike Xu, Tengyu Ma, and Simon S Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. *arXiv preprint arXiv:2102.04692*, 2021.

Donggyu Yun, Alexandre Proutiere, Sumyeong Ahn, Jinwoo Shin, and Yung Yi. Multi-armed bandit with additional observations. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2 (1):1–22, 2018.
A Lower bounds

A.1 Proof of Theorem 1

Theorem 1. Let $n^q_\infty(a) = \liminf_{T \to \infty} \frac{E_k[n^q_T(a)]}{T}$ for $a \in [K]$. If the bandit strategy is consistent w.r.t. $D$, then for any bandit instance $\nu \in D$, the following hold:

1. For any suboptimal arm $a \notin A_*(\nu)$,
   \[ n^q_\infty(a) \geq \frac{1}{K^+_\text{inf}(\nu_a, \mu^*, D)} . \]  

2. Assume that $a^*$ is the unique optimal arm and let $a \neq a^*$ be a suboptimal arm. Also, denote the maximal suboptimal reward by $\mu_s = \max_{a \neq a^*} \mu_a$. Then $n^q_\infty(a^*) \geq \frac{1}{K^+_\text{inf}(\nu_{a^*}, \mu^*, D)}$ and for any $\mu \in [\mu_s, \mu^*]$,
   \[ n^q_\infty(a) + n^q_\infty(a^*) \geq \frac{1}{\max\{K^+_\text{inf}(\nu_a, \mu, D), K^-\text{inf}(\nu_{a^*}, \mu, D)\}} . \]

3. Assume that there are at least two optimal arms. Moreover, assume that (1) for all optimal arms $a \in A_*(\nu)$, $K^+_\text{inf}(\nu_a, \mu^*, D) = 0$, or, alternatively, (2) there are at least two optimal arms for which $K^-\text{inf}(\nu_a, \mu^*, D) = 0$. Then $n^q_\infty(a) = \infty$ for some optimal arm $a \in A_*(\nu)$.

Proof. Parts of the proof rely on techniques from Theorem 1 in [Garivier et al., 2019]. First notice that for any $p, q \in [0, 1]$, it holds that

\[ \text{kl}(p, q) = p \ln \frac{q}{1-q} + (1-p) \ln \frac{1}{1-q} + p \ln p + (1-p) \ln(1-p) \] \[ \geq \ln(2) \] \[ \geq \max\{p \ln \frac{1}{q}, (1-p) \ln \frac{1}{1-q}\} - \ln 2 . \]  

(5)

**Proof of Part 1.** Let $a \notin A_*(\nu)$ be some suboptimal arm. Furthermore, let $\nu'$ be a modified bandit instance such that $\nu'_k = \nu_k$ for all $k \neq a$ and $\nu'_a \in D$ is some distribution with $E[\nu'_a] > \mu^*$ (if such a distribution does not exist, then $K^+_\text{inf}(\nu_a, \mu^*, D) = \infty$ and the bound trivially holds). Then, by Lemma 1 with $Z = \frac{n^q_T(a)}{T} \in (0, 1]$ and Equation (5), it holds that

\[ \mathbb{E}_{\nu'}[n^q_T(a)] \text{KL}(\nu_a, \nu'_a) \geq \text{kl}\left(\frac{\mathbb{E}_{\nu}[n^q_T(a)]}{T}, \frac{\mathbb{E}_{\nu'}[n^q_T(a)]}{T}\right) \] \[ \geq \left(1 - \frac{\mathbb{E}_{\nu}[n^q_T(a)]}{T}\right) \ln \frac{1}{1 - \frac{\mathbb{E}_{\nu'}[n^q_T(a)]}{T}} - \ln 2 . \]  

(6)

Since the bandit strategy is consistent and $a$ is suboptimal for instance $\nu'$, it holds that $\lim_{T \to \infty} \frac{\mathbb{E}_{\nu}[n^q_T(a)]}{T} = 0$. Moreover, $a$ is the unique optimal arm of bandit instance $\nu'$. Therefore, the consistency implies that for any $\alpha \in (0, 1]$ and for sufficiently large $T$,

\[ 1 - \frac{\mathbb{E}_{\nu}[n^q_T(a)]}{T} = T - \frac{\mathbb{E}_{\nu}[n^q_T(a)]}{T} = \frac{\sum_{a' \neq a} \mathbb{E}_{\nu}[n^q_T(a')]}{T} \leq \frac{T^\alpha}{T} = T^{\alpha-1} . \]

Combining both into eq. (6), we get that for any $\alpha \in (0, 1]$ and any $\nu'$ with $E[\nu'_a] > \mu^*$,

\[ \liminf_{T \to \infty} \frac{\mathbb{E}_{\nu'}[n^q_T(a)]}{\ln T} \geq \frac{1}{\text{KL}(\nu_a, \nu'_a)} \liminf_{T \to \infty} \frac{1}{\ln T} \left(\left(1 - \frac{\mathbb{E}_{\nu}[n^q_T(a)]}{T}\right) \ln \frac{1}{1 - \frac{\mathbb{E}_{\nu'}[n^q_T(a)]}{T}} - \ln 2\right) \] \[ \geq \frac{1}{\text{KL}(\nu_a, \nu'_a)} \liminf_{T \to \infty} \frac{1}{\ln T} \ln \frac{1}{T^{\alpha-1}} \] \[ = \frac{1 - \alpha}{\text{KL}(\nu_a, \nu'_a)} . \]
and since it holds for any \( \alpha \in (0, 1] \), we have that
\[
\liminf_{T \to \infty} \frac{E_\mu[n_T^\alpha(a)]} {\ln T} \geq \frac{1} {KL(\nu_a, \nu_a')}.
\]

By taking the supremum in the right-hand side over all distributions \( \nu_a' \in \mathcal{D} \) with \( E[\nu_a'] > \mu^* \), we conclude this part of the proof.

**Proof of Part 2.** Let \( \mu \in [\mu^*, \mu^*] \) and assume that there exist two distributions \( \nu_a', \nu_a'' \in \mathcal{D} \) such that \( E[\nu_a'] \geq \mu \) and \( E[\nu_a''] < \mu \) (otherwise, either \( K_{\text{inf}}^+(\nu_a, \mu, \mathcal{D}) = \infty \) or \( K_{\text{inf}}^-(\nu_a', \mu, \mathcal{D}) = \infty \) and the bound trivially holds). Also, define a new bandit instance \( \nu' \) for which \( \nu'_k = \nu_k \) for all \( k \neq a^* \) and arms \( a, a^* \) are distributed according to \( \nu_a' \) and \( \nu_a'' \), respectively. By Lemma 1 with \( Z = \frac{\nu(a^*)}{\nu(a)} \in [0, 1] \) and Equation (5), it holds that
\[
E_\nu[n_T^\alpha(a)]KL(\nu_a, \nu_a') + E_\nu[n_T^\alpha(a^*)]KL(\nu_{a^*}, \nu_a') \geq \text{kl} \left( \frac{E_\nu[n_T(a)]}{T}, \frac{E_\nu[n_T(a^*)]}{T} \right)
\]
\[
\geq \frac{E_\nu[n_T(a^*)]}{T} \ln \frac{1}{E_\nu[n_T(a^*)]} - 2.
\]

Since the bandit strategy is consistent and \( a^* \) is the unique optimal arm for instance \( \nu' \), for any \( a' \neq a^* \) we have that \( E_\nu'[n_T(a')]/T \to 0 \) and thus
\[
\lim_{T \to \infty} \frac{E_\nu[n_T(a^*)]}{T} = \lim_{T \to \infty} \frac{T - \sum_{a' \neq a^*} E_\nu[n_T(a')]}{T} = 1 - \sum_{a' \neq a^*} \lim_{T \to \infty} \frac{E_\nu[n_T(a')]}{T} = 1.
\]

Moreover, \( a^* \) is strictly suboptimal in bandit instance \( \nu' \). Therefore, the consistency implies that for any \( \alpha \in (0, 1] \) and for sufficiently large \( T \),
\[
\frac{E_\nu[n_T(a^*)]}{T} \leq \frac{T^\alpha}{T} = T^{\alpha-1}.
\]

Therefore, for any \( \alpha \in (0, 1] \), we can bound
\[
\liminf_{T \to \infty} \frac{1}{\ln T} \left( \frac{E_\nu[n_T(a^*)]}{T} \ln \frac{1}{E_\nu[n_T(a^*)]} - 2 \right) \geq \liminf_{T \to \infty} \frac{1}{\ln T} \ln \frac{1}{T^{\alpha-1}} = 1 - \alpha,
\]
and since it holds for any \( \alpha \in (0, 1] \), the same result holds for \( \alpha = 0 \).

For the l.h.s., we divide into two cases:

**Case I:** Letting \( \alpha \in \arg \max_{a' \neq a^*, \mu} \mu_{a'} \) be an arm such that \( \mu_a = \mu^* \) while choosing \( \mu = \mu^* \) and \( \nu_a' = \nu_a \), we get that \( KL(\nu_a, \nu_a') = 0 \). Furthermore, since \( \alpha \) is strictly suboptimal and \( E[\nu_a'] < \mu^* < \mu^* \), we have that \( KL(\nu_{a^*}, \nu_{a^*}) > 0 \). Dividing by it in (7) and combining with (8), we get that
\[
\liminf_{T \to \infty} \frac{E_\nu[n_T^\alpha(a)]}{\ln T} \geq \frac{1}{KL(\nu_{a^*}, \nu_{a^*})}.
\]

By taking the supremum in the right-hand side over all distributions \( \nu_a' \in \mathcal{D} \) with \( E[\nu_a'] < \mu^* \), we get the first desired result.

**Case II:** For any \( \mu \in [\mu^*, \mu^*] \), we apply Hölder’s inequality and bound the l.h.s. of Equation (7) by
\[
E_\nu[n_T^\alpha(a)]KL(\nu_a, \nu_a') + E_\nu[n_T^\alpha(a)]KL(\nu_{a^*}, \nu_a') \leq \left( E_\nu[n_T^\alpha(a)] + E_\nu[n_T^\alpha(a^*)] \right) \max\{KL(\nu_a, \nu_a'), KL(\nu_{a^*}, \nu_{a^*})\}.
\]

Importantly, notice that \( KL(\nu_{a^*}, \nu_{a^*}) > 0 \) (since \( E[\nu_{a^*}] < \mu \leq \mu^* \)) and we can divide by the maximum. Combining with (8), we get
\[
\liminf_{T \to \infty} \frac{E_\nu[n_T^\alpha(a)] + E_\nu[n_T^\alpha(a^*)]}{\ln T} \geq \frac{1}{\max\{KL(\nu_{a'}, \nu_a'), KL(\nu_{a^*}, \nu_{a^*})\}}.
\]
Taking the supremum over all distributions \( \nu'_a, \nu'_a \in D \) with expectations \( \mathbb{E}[\nu'_a] > \mu \) and \( \mathbb{E}[\nu'_a] < \mu \) leads to the second stated result and concludes this part of the proof.

We remark that a more general lower bound can be written without applying H"{o}lder’s inequality:

\[
\mathcal{K}^-_{\text{inf}}(\nu_a, \mu, D) n^3_\infty(a) + \mathcal{K}^+_{\text{inf}}(\nu_a, \mu, D) n^2_\infty(a^*) \geq 1
\]  

(9)

for any \( \mu \in [\mu^*, \mu^*] \), where we define \( 0 \cdot (+\infty) \geq 1 \). Notably, this definition makes sure that the bound holds whenever any of the quantities at its l.h.s. are infinite, so it is only needed to be proven when all quantities are finite. Then, starting from \( \square \), dividing by \( \ln T \), taking lim inf and using \( \square \), we get that

\[
\text{KL}(\nu_a, \nu'_a) n^3_\infty(a) + \text{KL}(\nu_a, \nu'_a) n^2_\infty(a^*) \geq 1
\]

Next, if both \( \mathcal{K}^-_{\text{inf}}(\nu_a, \mu, D), \mathcal{K}^+_{\text{inf}}(\nu_a, \mu, D) < +\infty \), we can take the infimum at the l.h.s. knowing that it is not over an empty set. Moreover, for \( n^3_\infty(a), n^2_\infty(a^*) < +\infty \), the inequality is preserved after the infimum, which leads to Equation \( \square \).

**Proof of Part 3.**

Assume that condition (1) holds. For any \( a \in A_+(\nu) \), define a bandit instance \( \nu' \) such that \( \nu'_k = \nu_k \) for all \( k \neq a \) and \( \nu'_a \in D \) is some distribution such that \( \mathbb{E}[\nu'_a] < \mu^* \) (such a distribution must exist, as otherwise, \( \mathcal{K}^-_{\text{inf}}(\nu_a, \mu^*, D) = +\infty \) and condition (1) does not hold). Then, by Lemma 1 with \( Z = \frac{\nu_a(a)}{\nu_a(a)} \in [0, 1] \) and Equation \( \square \), it holds that

\[
\sum_{a \in A_+(\nu)} \mathbb{E}_\nu[n^2_T(a)] \text{KL}(\nu_a, \nu'_a) \geq k \left( \frac{\mathbb{E}_\nu[n_T(a)]}{T}, \frac{\mathbb{E}_\nu'[n_T(a)]}{T} \right) \geq \frac{\mathbb{E}_\nu[n_T(a)]}{T} \ln \frac{1}{\mathbb{E}_\nu'[n_T(a)]} - \ln 2.
\]

As \( a \) is strictly suboptimal in bandit instance \( \nu' \), and since the bandit strategy is consistent, for any \( \alpha \in (0, 1) \) and for sufficiently large \( T \), it holds that

\[
\frac{\mathbb{E}_\nu'[n_T(a)]}{T} \leq T^{\alpha} T = T^{\alpha-1}.
\]

Then, for large enough \( T \), we have that

\[
\sum_{a \in A_+(\nu)} \mathbb{E}_\nu[n^2_T(a)] \text{KL}(\nu_a, \nu'_a) \geq \frac{\mathbb{E}_\nu[n_T(a)]}{T} \ln \frac{1}{T^{\alpha-1}} - \ln 2 = (1 - \alpha) \frac{\mathbb{E}_\nu[n_T(a)]}{T} \ln T - \ln 2.
\]

Since the number of arms is finite, for large enough \( T \), the inequality holds for all \( a \in A_+(\nu) \). Then, summing over all inequalities yields

\[
\sum_{a \in A_+(\nu)} \mathbb{E}_\nu[n^2_T(a)] \text{KL}(\nu_a, \nu'_a) \geq (1 - \alpha) \sum_{a \in A_+(\nu)} \mathbb{E}_\nu[n_T(a)] \ln T - K \ln 2.
\]

(10)

To further analyze the r.h.s. of the inequality, recall that the bandit strategy is consistent; therefore, the set of the optimal arm is sampled linearly, i.e.,

\[
\lim_{T \rightarrow \infty} \frac{\sum_{a \in A_+(\nu)} \mathbb{E}_\nu[n_T(a)]}{T} = \lim_{T \rightarrow \infty} T - \sum_{a' \notin A_+(\nu)} \mathbb{E}_\nu[n_T(a')]}{T} = 1 - \sum_{a' \notin A_+(\nu)} \lim_{T \rightarrow \infty} \frac{\mathbb{E}_\nu[n_T(a')]}{T} = 1,
\]

where the last equality is by the consistency, which implies that \( \mathbb{E}_\nu[n_T(a')]/T \rightarrow 0 \) for any \( a \notin A_+(\nu) \). For the l.h.s., we apply H"{o}lder’s inequality and get

\[
\sum_{a \in A_+(\nu)} \mathbb{E}_\nu[n^2_T(a)] \text{KL}(\nu_a, \nu'_a) \leq \max_{a \in A_+(\nu)} \text{KL}(\nu_a, \nu'_a) \sum_{a \in A_+(\nu)} \mathbb{E}_\nu[n^2_T(a)]
\]

Substituting both into Equation \( \square \), reorganizing and taking the limit, we get

\[
\lim_{T \rightarrow \infty} \frac{\sum_{a \in A_+(\nu)} \mathbb{E}_\nu[n^2_T(a)]}{\ln T} \geq (1 - \alpha) \frac{1}{\max_{a \in A_+(\nu)} \text{KL}(\nu_a, \nu'_a)}.
\]

15
As a result, for large enough $T$, to further analyze the r.h.s. of the inequality, notice that for

$$\sum_{a \in A, \nu \in D} E_{\nu}[n_T^a(a)] \geq \frac{1}{\max_{a \in A} \mathcal{K}^-_{inf}(\nu_a, \mu^*, D)} .$$

Importantly, this bound implies that if $\mathcal{K}^-_{inf}(\nu_a, \mu^*, D) = 0$ for all $a \in A, (\nu)$, then the r.h.s. is infinite and there exists at least one optimal arm $a \in A, (\nu)$ for which $n_T^a(a) = \lim inf_{T \to \infty} E_{\nu}[n_T^a(a)] = \infty$. Moreover, one can easily verify that the derivation did not actually require condition (1); therefore, this might serve as a lower bound when the condition does not hold. Finally, as in (9), if we define $0 \cdot (+\infty) = 1$, then using the exact same arguments, a more general version of the bound would be

$$\sum_{a \in A, (\nu)} \mathcal{K}^-_{inf}(\nu_a, \mu^*, D)n_T^a(a) \geq 1 .$$

(11)

Assume that condition (2) holds. For any $a \in A, (\nu)$, define a bandit instance $\nu'$ such that $\nu'_k = \nu_k$ for all $k \neq a$ and $\nu'_a \in D$ is some distribution such that $E[\nu'_a] > \mu^*$ (we will later take an infimum over all such distributions, and if such a distribution does not exist, then the value $\mathcal{K}^+_{inf}(\nu_a, \mu^*, D) = \infty$ will lead to a trivial bound). Then, by Lemma 4 with $Z = \frac{n_T^a(a)}{T} \in [0, 1]$ and Equation (11), it holds that

$$\mathbb{E}_{\nu}[n_T^a(a)] KL(\nu_a, \nu'_a) \geq kl\left(\frac{\mathbb{E}_{\nu}[n_T^a(a)]}{T}, \frac{\mathbb{E}_{\nu'}[n_T(a)]}{T}\right) \geq \left(1 - \frac{\mathbb{E}_{\nu}[n_T^a(a)]}{T}\right) \ln\frac{1}{1 - \frac{\mathbb{E}_{\nu'}[n_T(a)]}{T}} - \ln 2 .$$

As $a$ is the unique optimal arm for bandit instance $\nu'$, and since the bandit strategy is consistent, for any $a \in (0, 1)$ and for sufficiently large $T$ it holds that

$$1 - \frac{\mathbb{E}_{\nu'}[n_T(a)]}{T} = T - \frac{\mathbb{E}_{\nu'}[n_T(a)]}{T} = \sum_{a' \neq a} \mathbb{E}_{\nu'}[n_T(a')] \leq \frac{T^\alpha}{T} = T^{\alpha - 1} .$$

Then, for large enough $T$, we have that

$$\mathbb{E}_{\nu}[n_T^a(a)] KL(\nu_a, \nu'_a) \geq \left(1 - \frac{\mathbb{E}_{\nu}[n_T^a(a)]}{T}\right) \ln\frac{1}{T^{\alpha - 1}} - \ln 2 = (1 - \alpha)\left(1 - \frac{\mathbb{E}_{\nu}[n_T^a(a)]}{T}\right) \ln T - \ln 2 .$$

Next, let $a, b \in A, (\nu)$ be two optimal arms. For large enough $T$, the inequality holds for both arms, and summing over their respective inequalities yields

$$\mathbb{E}_{\nu}[n_T^a(a)] KL(\nu_a, \nu'_a) + \mathbb{E}_{\nu}[n_T^b(b)] KL(\nu_b, \nu'_b) \geq (1 - \alpha)\left(2 - \frac{\mathbb{E}_{\nu}[n_T^a(a)]}{T} - \frac{\mathbb{E}_{\nu}[n_T^b(b)]}{T}\right) \ln T - 2 \ln 2 .$$

(12)

To further analyze the r.h.s. of the inequality, notice that

$$2 - \frac{\mathbb{E}_{\nu}[n_T^a(a)]}{T} - \frac{\mathbb{E}_{\nu}[n_T^b(b)]}{T} \geq 2 - \sum_{a'} \mathbb{E}_{\nu}[n_T(a')] = 2 - \frac{T}{T} = 1 .$$

For the l.h.s., we bound using Hölder’s inequality:

$$\mathbb{E}_{\nu}[n_T^a(a)] KL(\nu_a, \nu'_a) + \mathbb{E}_{\nu}[n_T^b(b)] KL(\nu_b, \nu'_b) \leq \max\{KL(\nu_a, \nu'_a), KL(\nu_b, \nu'_b)\} (\mathbb{E}_{\nu}[n_T^a(a)] + \mathbb{E}_{\nu}[n_T^b(b)]) .$$
Substituting both into Equation (12), reorganizing and taking the limit, we get
\[ \lim_{T \to \infty} \frac{E_\nu[n^q_T(a)] + E_\nu[n^q_T(b)]}{\ln T} \geq (1 - \alpha) \frac{1}{\max \{KL(\nu_a, \mu_*), KL(\nu_b, \mu_*)\}}. \]

Taking the supremum over all distributions \( \nu'_a, \nu'_b \in \mathcal{D} \) such that \( E[\nu'_a] > \mu^*, E[\nu'_b] > \mu^* \) and noting that the bound holds for any \( \alpha \in (0, 1) \) leads to
\[ \lim_{T \to \infty} \frac{E_\nu[n^q_T(a)] + E_\nu[n^q_T(b)]}{\ln T} \geq \frac{1}{\max \{\mathcal{K}^+_{\inf}(\nu_a, \mu^*, \mathcal{D}), \mathcal{K}^+_{\inf}(\nu_b, \mu^*, \mathcal{D})\}}. \]

As we previously remarked, if one of the infimums is over an empty set, the r.h.s. equals zero and the bound trivially holds. Finally, as in the case of condition (1), if there exist two optimal arms \( a, b \in \mathcal{A}_*(\nu) \) for which \( \mathcal{K}^+_{\inf}(\nu_a, \mu^*, \mathcal{D}) = \mathcal{K}^+_{\inf}(\nu_b, \mu^*, \mathcal{D}) = 0 \), then the r.h.s. of this bound equals infinity, and for at least one of these arms \( \nu'_a \in \mathcal{D} \), it holds that \( n^q_\infty(a') = \lim_{T \to \infty} \frac{E[n^q_T(a')]}{\ln T} = \infty \). Furthermore, this bound can serve as a lower bound even when condition (2) does not hold. Finally, as in (9) and (11), if we define \( \theta \cdot (+\infty) \geq 1 \), a more general version of the bound would be
\[ \mathcal{K}^+_{\inf}(\nu_a, \mu^*, \mathcal{D})n^q_\infty(a) + \mathcal{K}^+_{\inf}(\nu_b, \mu^*, \mathcal{D})n^q_\infty(b) \geq 1. \]
A.2 Additional Intuition Behind the Lower Bound for multiple optimal arms

The following example illustrates the intuition behind the conditions for the third part of Theorem 1:

**Example 3.** Let \( x \in (0, 1) \). Define \( \mathcal{D}_1 \) as the set of all distributions with the discrete support \( \{0, 1\} \) and expectations in \([0, x]\). Also, define \( \mathcal{D}_2 \) as the set of all distributions with the discrete support \( \{x, 1\} \). Particularly, for any \( \nu \in \mathcal{D}_1 \), \( E[\nu] \leq x \), and for any \( \nu \in \mathcal{D}_2 \), \( E[\nu] \geq x \). Finally, let \( \mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \). Now let \( \nu \) be a bandit instance with arm distributions in \( \mathcal{D} \) such that \( \mu^* = x \), i.e., the optimal reward is \( x \). Then, the distribution of optimal arms is either \( \text{Ber}(x) \) (if \( \nu_{a^*} \in \mathcal{D}_1 \)) or outputs \( x \) w.p. 1 (if \( \nu_{a^*} \in \mathcal{D}_2 \)), and suboptimal arms are in \( \mathcal{D}_1 \). Consider the following instances:

1. **All optimal arms are in \( \mathcal{D}_1 \).** Recall that \( \Omega \left( \frac{\ln 1/\delta}{\epsilon^2} \right) \) samples are required to distinguish between \( \text{Ber}(x) \) and \( \text{Ber}(x - \epsilon) \) w.p. \( 1 - \delta \). Then, any fixed logarithmic budget cannot identify whether an arm is optimal or suboptimal with arbitrarily close mean \( x - \epsilon \), and at least one optimal arm must be queried super-logarithmically identify with high certainty that \( \mu^* = x \). Notice that in this case, \( K_{\inf}^-(\nu_a, x, \mathcal{D}) = 0 \) for all optimal arms.

2. **One optimal arm belongs to \( \mathcal{D}_2 \) and all other optimal arms are in \( \mathcal{D}_1 \).** In this case, an agent can identify the optimal arm in \( \mathcal{D}_2 \) by a single sample, since it outputs \( x \) w.p. 1 and this value is not supported by any distribution in \( \mathcal{D}_1 \). Similarly, an optimal arm can be related to \( \mathcal{D}_1 \) when outputs 0, which only requires \( O \left( \frac{\ln 1/\delta}{\ln 1/(1-\delta)} \right) \) samples with certainty \( 1 - \delta \). All suboptimal arms can be similarly identified. Then, with a logarithmic number of queries, agents can identify that only one arm belongs to \( \mathcal{D}_2 \), and as \( E[\nu] \leq E[\nu'] \) for any \( \nu \in \mathcal{D}_1 \) and \( \nu' \in \mathcal{D}_2 \), exploiting this arm is always optimal. Thus, logarithmic queries suffice.

3. **At least two optimal arms belong to \( \mathcal{D}_2 \).** In this case, agents must identify that the optimal mean \( \mu^* \) is not higher than \( x \), and as in the case where all optimal arms belong to \( \mathcal{D}_1 \), a logarithmic number of queries will not suffice, as it only allows identifying the mean of an arm with a fixed accuracy. Then, all optimal arms in \( \mathcal{D}_2 \), except for a single arm, must be sampled more than a logarithmic number of times to identify that their mean is exactly \( x \). In contrast, the remaining arm in \( \mathcal{D}_2 \) can be queried once to identify that it is in \( \mathcal{D}_2 \). Afterwards, as \( E[\nu] \geq x \) for any \( \nu \in \mathcal{D}_2 \) and \( \mu^* = x \), this arm can be safely exploited. This corresponds to the case where \( K_{\inf}^+(\nu_a, \mu^*, \mathcal{D}) = 0 \) for at least two optimal arms.

While the example illustrates that the conditions in the Theorem are sufficient, we actually believe that they are also necessary. In particular, when both conditions do not hold, we believe that an action-elimination algorithm that uses KL-based confidence interval [Garivier et al. 2018] should allow logarithmically querying the optimal arms. To do so, the algorithm will have to prioritize the optimal arm that might have similar distributions with higher expectations, if exists (e.g., the optimal arm in \( \mathcal{D}_2 \) in the second part of the example). However, this is not a formal proof, which we leave for future work.
A.3 Proofs of Proposition 2 and Corollary 3

Proposition 2. For any bandit instance \( \nu \in D \), any strategy that is better than uniform on \( D \), any arm \( a \notin A_*(\nu) \) and any \( T \geq 1 \), it holds that

\[
E_\nu[n_T(a)] \geq \frac{T}{K} \left( 1 - \sqrt{2 \min \{ KE_\nu[n_T^q(a)] , T \} K_{\inf}^+(\nu_a, \nu^*, D) } \right).
\]

If \( K_{\inf}^+(\nu_a, \nu^*, D) = \infty \), we define the r.h.s. to be zero. Specifically, if \( K_{\inf}^+(\nu_a, \nu^*, D) < \infty \) and \( E_\nu[n_T^q(a)] \leq \frac{1}{8 K K_{\inf}^+(\nu_a, \nu^*, D)} \), then \( E_\nu[n_T(a)] \geq T/(2K) \).

Proof. We follow the proof of Theorem 2 of Garivier et al. (2019). Let \( a \notin A_*(\nu) \) and let \( \nu' \) be some modified problem such that \( \nu'_k = \nu_k \) for all \( k \neq a \) and \( \nu'_a \in D \) is such that \( E[\nu'_a] > \mu^* \) (if no such distribution exists, then \( K_{\inf}^+(\nu_a, \nu^*, D) = \infty \) and the r.h.s. of the lower bound is defined as 0, so the bound trivially holds). Furthermore, notice that the desired lower bound is always smaller than \( \frac{T}{K} \); therefore, if \( E_\nu[n_T(a)] \geq \frac{T}{K} \) then the bound holds. Thus, we can assume for the rest of the proof that \( E_\nu[n_T(a)] < \frac{T}{K} \). Finally, since the strategy is better than uniform on \( D \) and \( a \) is the unique optimal arm in \( \nu' \), it holds that \( E_\nu[n_T(a)] \geq \frac{T}{K} \).

Then, by Lemma 1 with \( Z = \frac{n_T(a)}{T} \in [0, 1] \), we have that

\[
E_\nu[n_T^q(a)] KL(\nu_a, \nu'_a) \geq k l \left( \frac{E_\nu[n_T^q(a)]}{T} , \frac{E_\nu[n_T(a)]}{T} \right) \geq k l \left( \frac{E_\nu[n_T^q(a)]}{T} , \frac{1}{K} \right) ,
\]

where the last inequality is since \( \frac{E_\nu[n_T(a)]}{T} < \frac{T}{K} \), \( E_\nu[n_T^q(a)] \geq \frac{1}{K} \) and for \( p \leq q \), the function \( k l(p, q') \) is increasing in \( q' \in [p, q] \). Next, recall that by the local refinement of Pinsker’s inequality (e.g., Lemma 6 in Garivier et al. 2019), for any \( 0 \leq p < q \leq 1 \), we have that

\[
kl(p, q) \geq \frac{1}{2q} (q - p)^2
\]

Substituting into Equation (14), we get

\[
E_\nu[n_T^q(a)] KL(\nu_a, \nu'_a) \geq \frac{K}{2} \left( \frac{1}{K} - \frac{E_\nu[n_T(a)]}{T} \right)^2 ,
\]

which can be reorganized (using \( \frac{1}{K} \geq \frac{E_\nu[n_T(a)]}{T} \)) to

\[
E_\nu[n_T(a)] \geq \frac{T}{K} \left( 1 - \sqrt{2 K E_\nu[n_T^q(a)] KL(\nu_a, \nu'_a)} \right) .
\]

Moreover, since \( E_\nu[n_T(a)] \leq \frac{T}{K} \), we also have that \( E_\nu[n_T^q(a)] \leq \frac{T}{K} \). This leads to a bound of

\[
E_\nu[n_T(a)] \geq \frac{T}{K} \left( 1 - \sqrt{2T \cdot KL(\nu_a, \nu'_a)} \right) .
\]

Taking the maximum between both previous bounds results with

\[
E_\nu[n_T(a)] \geq \frac{T}{K} \left( 1 - \sqrt{2 \min \{ KE_\nu[n_T^q(a)] , T \} KL(\nu_a, \nu'_a)} \right) ,
\]

and taking the supremum over all distributions \( \nu'_a \in D \) such that \( E[\nu'_a] > \mu^* \) leads to the desired result. In particular, notice that the result when \( K_{\inf}^+(\nu_a, \nu^*, D) < \infty \) and \( E_\nu[n_T^q(a)] \leq \frac{1}{8 K K_{\inf}^+(\nu_a, \nu^*, D)} \) directly follows from the general bound. \( \square \)
Corollary 3. Assume that for any \( t \geq 1 \) and \( a \in [K] \), \( E_u[n_T^a(t)] \leq B_a(t) \) for some positive nondecreasing budget \( \{B_a(t)\}_{t \geq 1} \). Also let \( B_a^{-1}(x) = \sup\{t \in \mathbb{N} : B_a(t) \leq x\} \) for \( x \geq B(1) \) and otherwise \( B_a^{-1}(x) = 0 \). Then, for any instance \( u \in \mathcal{D} \) and any better-than-uniform bandit strategy on \( \mathcal{D} \),

\[
\forall T \geq \max_{a \notin \mathcal{A}_u(u)} B_a^{-1}\left(\frac{1}{8KK^{+}_{\inf}(\nu_a, \mu^*, \mathcal{D})}\right), \quad \text{Reg}(T) \geq \sum_{a \notin \mathcal{A}_u(u)} \frac{\Delta_a}{2K} B_a^{-1}\left(\frac{1}{8K K^{+}_{\inf}(\nu_a, \mu^*, \mathcal{D})}\right).
\]

Proof. First, notice that since we defined \( B_a^{-1}(x) = 0 \) when \( B(1) > x \), then \( B_a^{-1}(x) \) always exists, but might be equal to +\( \infty \) if \( B(t) \leq x \) for all \( t \in \mathbb{N} \). On the other hand, if \( B_a^{-1}(x) = +\infty \), then the bound trivially holds (as there is no \( T \) for which the result must hold). Thus, w.l.o.g., we assume that \( \sup\{t \in \mathbb{N} : B_a(t) \leq x\} < +\infty \), and thus \( B_a^{-1}(x) = \max\{t \in \mathbb{N} : B_a(t) \leq x\} \).

Moreover, by definition, for any suboptimal arm with \( \frac{1}{8KK^{+}_{\inf}(\nu_a, \mu^*, \mathcal{D})} < B(1) \), it holds that \( B_a^{-1}\left(\frac{1}{8KK^{+}_{\inf}(\nu_a, \mu^*, \mathcal{D})}\right) = 0 \), and such arms do not affect both the time constraint nor the regret bound. Therefore, we also assume w.l.o.g. that \( \frac{1}{8KK^{+}_{\inf}(\nu_a, \mu^*, \mathcal{D})} \geq B(1) \) for all suboptimal arms. Particularly, since \( B(1) > 0 \), this condition also implies that \( K^{+}_{\inf}(\nu_a, \mu^*, \mathcal{D}) \) is finite for all \( a \notin \mathcal{A}_u(u) \).

Denote \( T_a = B_a^{-1}\left(\frac{1}{8KK^{+}_{\inf}(\nu_a, \mu^*, \mathcal{D})}\right) \). Then, by definition, we have that

\[
E_u[n_{T_a}(a)] \leq B_a(T_a) \leq \frac{1}{8K K^{+}_{\inf}(\nu_a, \mu^*, \mathcal{D})}.
\]

In turn, Proposition \[2\] leads to the bound \( E_u[n_{T_a}(a)] \geq \frac{T_a}{2K} \). Finally, as \( E_u[n_T(a)] \) is nondecreasing in \( T \), for any \( T \geq \max_{a \notin \mathcal{A}_u(u)} T_a \), we have that

\[
\text{Reg}(T) = \sum_{a \notin \mathcal{A}_u(u)} \Delta_a \mathbb{E}_u[n_T(a)]
\geq \sum_{a \notin \mathcal{A}_u(u)} \Delta_a \mathbb{E}_u[n_{T_a}(a)]
\geq \sum_{a \notin \mathcal{A}_u(u)} \Delta_a T_a
\geq \sum_{a \notin \mathcal{A}_u(u)} \frac{\Delta_a}{2K} B_a^{-1}\left(\frac{1}{8K K^{+}_{\inf}(\nu_a, \mu^*, \mathcal{D})}\right).
\]

\( \square \)
B Upper Bounds

In this part of the appendix, we prove the upper bounds of Section 4. In particular, to make the proof as general as possible, we prove the upper bounds for any confidence intervals that follow some mild regularity assumptions. First recall that we denoted the history of the bandit process by $I_t = (U_0, a_1, q_1, R_1 \cdot q_1, U_1, \ldots, a_t, q_t, R_t \cdot q_t, U_t)$ (see Section 3), and we further define $F_t = \sigma(I_t)$. Then, we define regular confidence intervals as follows:

**Definition 3.** Let $T \geq 1$ and let $\{[LCB_t(a), UCB_t(a)]\}_{t \in [T], a \in [K]}$ be a sequence of confidence intervals such that $LCB_t(a), UCB_t(a)$ are predictable w.r.t. $F_t$ for any $t \in [T]$ and $a \in [K]$. Then, the confidence intervals are called regular if there exists a sequence of events $\{G_t \in F_{t-1}\}_{t \in [T]}$ (‘good events’) such that regardless of the bandit strategy, the following hold:

1. For any $t \in [T]$ and $a \in [K]$, it holds that $CI_t(a) \equiv UCB_t(a) - LCB_t(a) \geq 0$.
2. For any $t \in [T]$, if $G_t$ holds, then $\mu_a \in [LCB_t(a), UCB_t(a)]$ for all $a \in [K]$.
3. The failure probabilities are bounded by $\sum_{t=1}^{T} \Pr\{\overline{G}_t\} \leq C(T)$.
4. For any $a \in [K]$, there exists a function $N^g_a : [T] \times [0, \Delta_a] \mapsto [0, \infty)$ such that if $G_t$ holds, then for any $N^g_a(t, \mu) < n_{t-1}^g(a) \leq t - 1$, it holds that $CI_t(a) < \mu$. Moreover, $N^g_a(t, \mu) \leq t - 1$ for all $a, t$ and $\mu$ and we define $N^g_a(0, \mu) = t - 1$.
5. There exists a function $N^{as}_a : [T] \times [0, \infty) \mapsto [0, \infty]$ such that for any $a \in [K]$ and $N^{as}_a(t, \mu) < n_{t-1}^{as}(a) \leq t - 1$, it holds that $CI_t(a) < \mu$. Moreover, $N^{as}_a(t, \mu) \leq t - 1$ for all $a, t$ and $\mu$ and we define $N^{as}_a(0, \mu) = t - 1$.

Finally, w.l.o.g., we assume that for any $t \geq 1$, $a \in [K]$ and $\mu \in [0, \Delta_a]$, it holds that $N^g_a(t, \mu) \leq N^{as}_a(t, \mu)$, as by definition, we can always replace $N^g_a(t, \mu)$ by $\min\{N^g_a(t, \mu), N^{as}_a(t, \mu)\}$. Equivalently, any bound dependence in $N^g_a(t, \mu)$ can be replaced by $\min\{N^g_a(t, \mu), N^{as}_a(t, \mu)\}$

Each of the conditions is a reasonable requirement from confidence intervals. The first condition requires that $[LCB_t(a), UCB_t(a)]$ will be a nonempty interval. The second condition requires it to contain the true mean under some good event, while the third makes sure that the good events hold with sufficiently high probability. The last two conditions characterize the quantities that will affect the performance when using the confidence intervals: condition four quantifies the number of samples $N^g_a(t, \mu)$ required for an arm $a$ to be well-concentrated (up to a confidence level $\mu$) at time $t$ when the good event holds. Importantly, it will determine in-expectation regret and querying bounds. We emphasize that $N^g_a$ might depend on the specific arm distribution $\nu_a$ (e.g., be variance-dependent), but cannot depend on any random quantity. The last condition quantifies the number of samples $N^{as}_a(t, \mu)$ required for the confidence intervals to be smaller than $\mu$ at time $t$ regardless of the good event. In particular, this condition must hold for all arms and will determine the almost-sure querying guarantees. In practice, all these requirements are extremely mild and hold for most standard confidence intervals, including Hoeffding-based confidence intervals (as in the main paper) and Bernstein-based confidence intervals. We refer the readers to Appendix B.3 for the regularity proofs of these confidence intervals. We now state a more general version of Theorem 2 that characterizes the performance of Algorithm 1 when used with regular confidence intervals:

**Theorem 3.** Let $T \geq 1$ and let $\{\epsilon(t)\}_{t \in [T]}$ be some nonnegative sequence. Also, let $L_\epsilon(T, \Delta) = \sum_{t=1}^{T} 1\{\epsilon(t) \geq \Delta\}$ be the number of times $\epsilon(t)$ exceeds a confidence-level $\Delta \geq 0$ until $T$ and define

$$\bar{N}^g_a(T, \mu) = \max_{t \in [T]} N^g_a(t, \max\{\mu, \epsilon(t)\}), \quad \bar{N}^{as}(T) = \max_{t \in [T]} N^{as}(t, \epsilon(t)) \, .$$

Then, when applying Algorithm 1 with regular confidence intervals, the following hold:

1. For all $a \in [K]$, it holds that $n^g_a(a) \leq \bar{N}^{as}(T) + 1$. 

21
2. If there are multiple optimal arms \( |\mathcal{A}_*| > 1 \), then

\[
\text{Reg}(T) \leq \sum_{a \notin \mathcal{A}_*} \Delta_a \left( \hat{N}_a^g(T, \Delta_a) + L_{\epsilon}(T, \Delta_a) \right) + (K + C(T)) \Delta_{\text{max}},
\]

\[
\mathbb{E}[B^q(T)] \leq \sum_{a=1}^{K} \hat{N}_a^g(T, \Delta_a) + (K + C(T)) \Delta_{\text{max}}.
\]

3. If the optimal arm \( a^* \) is unique, then

\[
\text{Reg}(T) \leq \sum_{a \neq a^*} \Delta_a \left( \hat{N}_a^g \left( T, \frac{\Delta_a}{2} \right) + L_{\epsilon}(T, \Delta_a) \right) + (K + C(T)) \Delta_{\text{max}},
\]

\[
\mathbb{E}[B^q(T)] \leq \sum_{a \neq a^*} \hat{N}_a^g \left( T, \frac{\Delta_a}{2} \right) + \hat{N}_{a^*}^g \left( T, \frac{\Delta_{\text{min}}}{2} \right) + (K + C(T)).
\]

Notably, the bounds hold for any sequence of nonnegative confidence levels \( \{\epsilon(t)\}_{t \geq 1} \). This stands in sharp contrast to the results of Efroni et al. [2021], which assumed that the budget is nondecreasing, a requirement that limits the valid sequences of \( \epsilon(t) \). However, the cost for this generalization is the maximization over \( t \in [T] \) in \( \hat{N}_a^g(T, \mu) \) and \( \hat{N}_{a^*}^g(T) \). In particular, the maximization ensures that regardless of \( t \), these quantities upper bound the number of samples required for the confidence intervals to be small. Nonetheless, for reasonable choices of \( \epsilon(t) \), \( \hat{N}_a^g(T, \mu) \) and \( \hat{N}_{a^*}^g(T) \) can be easily calculated. We refer the readers to Lemma 7 and Lemma 8, where we present natural choices of \( \epsilon \) for Hoeffding and Bernstein confidence intervals, respectively, and explicitly bound \( \hat{N}_a^g \) and \( \hat{N}_{a^*}^g \). Specifically, one favorable choice is to require \( \epsilon(t) \) to be nonincreasing. Then, bounding \( L_{\epsilon} \) by \( L_{\epsilon}(T, \Delta) \leq \epsilon^{-1}(\Delta) \triangleq \sup\{t \geq 1 : \epsilon(t) \geq \Delta\} \) is (asymptotically) tight and much easier to understand and compute. Finally, we remark that for ease of writing, in both aforementioned lemmas, \( B(t) \) is the per-arm querying budget. Then, results for total querying budget (as we immediately discuss) are simply obtained by replacing \( B(t) \rightarrow B(t)/K \).

When applied with Hoeffding-based confidence intervals, the theorem reduces to the results in Theorem 2 and in particular, choosing \( \epsilon(t) = \sqrt{\frac{6K \ln t}{B(t)}} \) for a nondecreasing positive budget \( B(t) \) also provides budget guarantees, namely, \( B^q(T) \leq B(T) + K \). Interestingly, a similar choice with Bernstein-based confidence intervals (explicitly, \( \epsilon(t) = \sqrt{\frac{6K \ln t}{B(t)-K}} + 14 \frac{K \ln t}{B(t)-K} \) for \( B(t) > K \)) provides regret bounds that depend on the variance of arms \( \{V_a\}_{a \in [K]} \) and expected per-arm querying bounds of \( O(4V_\text{a}B(t)/K) \). Importantly, since \( V_a \leq \frac{1}{4} \), this improves the Hoeffding-based bounds and might be dramatically lower when the variance is low.
B.1 Proof of Theorem 3

Before proving the theorem, we start by proving a few key properties on the querying mechanism and played actions of Algorithm 1.

**Lemma 5.** Let \( t \in \{K + 1, \ldots, T\} \) and assume that Algorithm 1 is run with regular confidence intervals and \( \epsilon(t) \geq 0 \). Then, the following hold: (i) If \( q_t = 1 \), then \( CI_t(a_t) > \epsilon(t) \). (ii) If the good event \( \mathcal{G}_t \) holds and \( q_t = 0 \), then \( \Delta_n \leq \epsilon(t) \).

**Proof.**

**Part (i).** When \( q_t = 1 \), recall that \( a_t = c_t \in \{l_t, u_t\} \). We divide into two cases. If \( a_t = l_t \), then a necessary condition for querying is that

\[
\epsilon(t) < UCB_t(c_t) - LCB_t(l_t) = UCB_t(l_t) - LCB_t(l_t) = CI_t(l_t) ,
\]

as required. Otherwise, \( a_t = u_t \). Then, by the definition of \( l_t \), we have that \( LCB_t(u_t) \leq LCB_t(l_t) \) and the querying condition implies that

\[
\epsilon(t) < UCB_t(c_t) - LCB_t(l_t) = UCB_t(u_t) - LCB_t(l_t) \leq UCB_t(u_t) - LCB_t(u_t) = CI_t(u_t) ,
\]

which concludes this part of the proof.

**Part (ii).** When \( q_t = 0 \), we have that \( a_t = l_t \); therefore, we need to show that \( \Delta_t \leq \epsilon(t) \). For \( q_t = 0 \) to hold, at least one of the following two options must occur:

1. **First option:** \( UCB_t(u_t) \leq LCB_t(l_t) \), which implies that \( UCB_t(a) \leq LCB_t(l_t) \) for all \( a \neq l_t \). In particular, since the good event \( \mathcal{G}_t \) holds, we have that \( LCB_t(l_t) \leq \mu_{l_t} \) and \( UCB_t(a) \geq \mu_a \) for all \( a \neq l_t \), and thus, \( \mu_a \leq \mu_{l_t} \) for all \( a \neq l_t \). Therefore, \( l_t \) is an optimal arm and \( \Delta_t = 0 \leq \epsilon(t) \).

2. **Second option:** \( UCB_t(c_t) - LCB_t(l_t) \leq \epsilon(t) \). Assume w.l.o.g. that \( l_t \) is not an optimal arm, as otherwise \( \Delta_t = 0 \) and the claim naturally holds, and let \( a^* \neq l_t \) be an optimal arm. Specifically, under the good event \( \mathcal{G}_t \), it holds that

\[
UCB_t(u_t) = \max_{a \neq l_t} UCB_t(a) \geq UCB_t(a^*) \geq \mu_{a^*} = \mu^* .
\]

We now divide into the cases where \( c_t = u_t \) and \( c_t = l_t \). If \( c_t = u_t \), we use the fact that under \( \mathcal{G}_t \), \( LCB_t(l_t) \leq \mu_{l_t} \), and we get that

\[
\Delta_t = \mu^* - \mu_{l_t} \leq UCB_t(u_t) - LCB_t(l_t) \leq \epsilon(t) ,
\]

where the last inequality is by the assumption that \( UCB_t(c_t) - LCB_t(l_t) \leq \epsilon(t) \) for \( c_t = u_t \). On the other hand, if \( c_t = l_t \), this assumption is equivalent to \( CI_t(l_t) \leq \epsilon(t) \). In turn, since \( l_t = c_t \in \arg \max_{a \in \{u_t, l_t\}} CI_t(a) \), it also implies that \( CI_t(u_t) \leq \epsilon(t) \). Finally, under the good event, and since \( l_t \in \arg \max_{a} LCB_t(a) \), we have that

\[
\mu_{l_t} \geq LCB_t(l_t) \geq LCB_t(u_t) \]

Then, recalling that \( UCB_t(u_t) \geq \mu^* \) and \( CI_t(u_t) \leq \epsilon(t) \) leads to the result since

\[
\Delta_t = \mu^* - \mu_{l_t} \leq UCB_t(u_t) - LCB_t(u_t) = CI_t(u_t) \leq \epsilon(t) .
\]

\( \square \)
Given this lemma, we can now prove Theorem 3.

**Proof.** We remark that at the first K rounds, each arm is played and queried once. Then, it can be verified that all bounds hold for T ≤ K. Thus, throughout the proof, we assume w.l.o.g. that T > K.

**Almost-sure querying bound:** Assume in contradiction that \( n^q_T(a) > \bar{N}^{as}(T) + 1 \) for some \( a \in [K] \). Therefore, there exists a time index \( t \in \{K + 1, \ldots, T\} \) such that action \( a \) was queried (\( a_t = a \) and \( q_t = 1 \)) and \( n^q_{t-1}(a) > \bar{N}^{as}(T) \geq N^{as}(t,\epsilon(t)) \), where the last inequality is by the definition of \( N^{as} \). In particular, since \( n^q_{t-1}(a) \leq t - 1 \), this condition cannot hold if \( N^{as}(t) = t - 1 \). Therefore, \( N^{as}(t,\epsilon(t)) < t - 1 \), and by the definition of \( N^{as} \), we have that \( CI_t(a) < \epsilon(t) \). This come in contradiction to the fact that \( a_t = a \) was queried, since part (i) of Lemma 5 implies that \( CI_t(a) = CI_t(a_t) > \epsilon(t) \). This proves that \( n^q_T(a) \leq \bar{N}^{as}(T) + 1 \) for all \( a \in [K] \) and concludes the first part of the proof.

**Count decomposition:** To derive the expected regret and querying bounds, we start with a general decomposition that will be relevant for all the required results. Specifically, the number of plays of any arm \( a \) under the good event \( G_t \) (defined in Definition 3) can be bounded as follows:

\[
\sum_{t=1}^{T} \{ G_t, a_t = a \} = \sum_{t=1}^{T} \{ G_t, a_t = a, q_t = 0 \} + \sum_{t=1}^{T} \{ G_t, a_t = a, q_t = 1 \}.
\]  

(15)

By Lemma 5(part (i)), the first term can be bounded by

\[
\sum_{t=1}^{T} \{ G_t, a_t = a, q_t = 0 \} = \sum_{t=1}^{T} \{ G_t, l_t = a, q_t = 0 \} \leq \sum_{t=1}^{T} \{ G_t, \Delta_a \leq \epsilon(t) \} \leq L\epsilon(T, \Delta_a),
\]

(16)

where the last relation is by the definition of \( L\epsilon \). For the second term of (15), let \( \Delta \geq 0 \) be some parameter that will be determined later. Then, under the good event, we bound

\[
\sum_{t=1}^{T} \{ G_t, a_t = a, q_t = 1 \} = \sum_{t=1}^{T} \left\{ G_t, a_t = a, q_t = 1, n^q_{t-1}(a) \leq \bar{N}^q_a(T, \Delta) \right\} \leq \bar{N}^q_a(T, \Delta) + 1,
\]

which leads to a total bound of

\[
\sum_{t=1}^{T} \{ G_t, a_t = a \} \leq \bar{N}^q_a(T, \Delta) + 1 + L\epsilon(T, \Delta_a).
\]

(18)

**Proving the bound of Equation (17).** The bound on the first term holds since \( n^q_T(a) \) starts from zero and increases by one every time action \( a \) was played and queried. We now show that depending on the assumptions and specific arms, \( \Delta \) can be chosen such that under \( G_t \), the events (s) of the second term cannot occur. This is trivially true if \( t \leq K \), since each arm \( a \) is queried at time \( t = a \) with \( n^q_{t-1}(a) = 0 \) and \( N^{\Delta \times \epsilon}_{a}(a) \geq 0 \). Therefore, w.l.o.g., we focus on \( t > K \). Moreover, by the definition of \( \bar{N}^q_a \), we have that \( \bar{N}^q_a(T, \Delta) \geq \bar{N}^q_a(t, \max(\Delta, \epsilon(t))) \). Then, it suffices to show that under \( G_t \), action \( a \) cannot be queried if \( n^q_{t-1}(a) > \bar{N}^q_a(t, \max(\Delta, \epsilon(t))) \).

To show this, first recall that \( n^q_{t-1}(a) \leq t - 1 \); thus, if \( \bar{N}^q_a(T, \max(\Delta, \epsilon(t))) = t - 1 \), this condition can never hold. Otherwise, \( \bar{N}^q_a(T, \max(\Delta, \epsilon(t))) < t - 1 \), and by the regularity of the confidence interval, when \( G_t \) holds and \( n^q_{t-1}(a) > \bar{N}^q_a(T, \max(\Delta, \epsilon(t))) \) we have that

\[
CI_t(a_t) = CI_t(a) < \max(\Delta, \epsilon(t)).
\]
Importantly, if \( \max\{\Delta, \epsilon(t)\} = \epsilon(t) \), then the condition of \( CI_t(a_t) < \epsilon(t) \) implies that an action cannot be queried (namely, \( q_t \neq 1 \)), by the first part of Lemma 5. Therefore, w.l.o.g., we assume that \( \max\{\Delta, \epsilon(t)\} = \Delta \) and prove by contradiction that for the right choice of \( \Delta \), \( a_t = a \) cannot be queried (under \( G_t \)) if \( CI_t(a) < \Delta \). This will imply that all indicators in (\( * \)) are equal to zero and will conclude the proof of Equation (17). To do so, divide into the cases where \( a \) is optimal or suboptimal and problems with unique or multiple optimal arms.

In all cases, assume in contradiction that \( a \) is queried and recall that it implies that \( a = a_t = c_t \in \{ l_t, u_t \} \).

(i) \( a \) is an optimal arm and \( \Delta = 0 \).

By the regularity of the confidence intervals, \( CI_t(a_t) \geq 0 \). Therefore, the condition \( CI_t(a_t) < 0 = \Delta \) can never hold.

(ii) \( a \) is strictly suboptimal, \( a_t = a = l_t \) and \( \Delta \leq \Delta_a \).

Let \( a^* \) be any optimal arm \( a^* \) (which is different than \( l_t = a \) since it is suboptimal). Then, the good event implies that \( UCB_t(a^*) \geq \mu^* \), and thus

\[
UCB_t(u_t) = \max_{a' \neq l_t} UCB_t(a') \geq UCB_t(a^*) \geq \mu^* .
\]

On the other hand, the good event also implies that \( LCB_t(l_t) \leq \mu_{l_t} = \mu_a \), and by definition of \( l_t \), it holds that \( LCB_t(u_t) \leq LCB_t(l_t) \leq \mu_a \). Combining both inequalities, we get that

\[
CI_t(u_t) = UCB_t(u_t) - LCB_t(u_t) \geq \mu^* - \mu_a = \Delta_a .
\]

However, since \( a_t = l_t \) was played and queried, by the definition of \( c_t \), it must hold that \( CI_t(l_t) \geq CI_t(u_t) \geq \Delta_a \), in contradiction to the fact that \( CI_t(l_t) = CI_t(a) < \Delta \leq \Delta_a \).

(iii) \( a \) is strictly suboptimal, \( a_t = a = u_t \) and \( \Delta \leq \Delta_u \): multiple optimal arms.

Since there are at least two optimal arms, there exists at least one optimal arm \( a^* \) such that \( l_t \neq a^* \). Specifically, under the good event, we have that \( UCB_t(a^*) \geq \mu^* \), and then

\[
UCB_t(u_t) = \max_{a' \neq l_t} UCB_t(a') \geq UCB_t(a^*) = \mu^* .
\]

Moreover, under \( G_t \), we have that \( LCB_t(u_t) \leq \mu_{u_t} \), and combining both leads to

\[
CI_t(u_t) = UCB_t(u_t) - LCB_t(u_t) \geq \mu^* - \mu_{u_t} = \Delta_{u_t} ,
\]

in contradiction to the fact that \( u_t = a \) and \( CI_t(a) < \Delta \leq \Delta_u \).

(iv) \( a \) is strictly suboptimal, \( a_t = a = u_t \) and \( \Delta \leq \Delta_u / 2 \): unique optimal arm \( a^* \).

Since \( a_t = c_t = u_t \), it holds that \( CI_t(u_t) < \Delta \leq \frac{\Delta_u}{2} \). In particular, \( G_t \) implies that

\[
UCB_t(u_t) = LCB_t(u_t) + CI_t(u_t) \leq \mu_{u_t} + CI_t(u_t) < \mu_{u_t} + \frac{\Delta_{u_t}}{2} = \mu^* - \frac{\Delta_{u_t}}{2} .
\]

On the other hand, under \( G_t \), it holds that \( UCB_t(a^*) \geq \mu^* > UCB_t(u_t) \). Then, the scenario of \( a = u_t \) can only happen if \( l_t = a^* \) (as \( u_t \) maximizes the UCB only on actions different than \( l_t \)), and we get that

\[
LCB_t(l_t) = UCB_t(l_t) - CI_t(l_t) = UCB_t(a^*) - CI_t(l_t) \geq \mu^* - CI_t(l_t) .
\]

Finally, since \( a_t = c_t = u_t \), and by the definition of \( c_t \), it holds that \( CI_t(l_t) \leq CI_t(u_t) < \frac{\Delta_{u_t}}{2} \), and we have that

\[
LCB_t(l_t) > \mu^* - \frac{\Delta_{u_t}}{2} > UCB_t(u_t) ,
\]

in contradiction to the querying rule.
(v) \(a = a^*\) is a unique optimal arm and \(\Delta \leq \Delta_{\min}/2\).

In this case, by the good event and the requirement that \(CI_t(a) = CI_t(a^*) < \Delta\),

\[
LCB_t(a^*) = UCB_t(a^*) - CI_t(a^*) \geq \mu^* - CI_t(a^*) > \mu^* - \Delta \geq \mu^* - \Delta_{\min}/2.
\]

Specifically, \(LCB_t(a^*) > \mu_a \geq LCB_t(a)\) for all \(a \neq a^*\) and thus \(l_t = a^*\). Furthermore, by the uniqueness of the optimal arm, \(u_t \in \text{arg max}_{a \neq a^*} UCB_t(a)\) is a strictly suboptimal arm. Therefore, under \(G_t\), we have that

\[
UCB_t(u_t) = LCB_t(u_t) + CI_t(u_t) \leq \mu_{u_t} + CI_t(u_t) \leq \mu^* - \Delta_{\min} + CI_t(u_t)
\]

where the last inequality is since \(u_t\) is a suboptimal arm (and so, \(\Delta_{u_t} \geq \Delta_{\min}\)). Finally, recall that \(q_t = 1\) only if \(UCB_t(u_t) > LCB_t(l_t)\). However, the previous inequalities imply that this condition can only hold if \(CI_t(u_t) > \frac{\Delta_{\min}}{2} > CI_t(a^*) = CI_t(l_t)\), and then \(c_t = \text{arg max}_{a \in \{u_t, l_t\}} CI_t(a) = u_t\). Thus, if an arm is indeed queried, it is a strictly suboptimal arm, in contradiction to the requirement that \(a = a^*\) is queried.

Overall, when there are multiple optimal arms, cases (i)-(iii) all hold for the choice \(\Delta = \Delta_a\). When the optimal arm is unique, cases (ii), (iv) and (v) hold with \(\Delta = \Delta_a/2\) for suboptimal arms and \(\Delta = \Delta_{\min}/2\) for the optimal arm.

**Regret and queries bounds.** We now show how the bounds of (17) and (18) can be used to derive the desired regret and querying bounds. To bound the expected regret, we use Equation (18) with \(\Delta = \mu_a\) for all suboptimal arms \(a \notin \mathcal{A}_a\), where \(\mu_a = \Delta_a\) if there are multiple optimal arms and \(\mu_a = \Delta_a/2\) if the optimal arm is unique. Doing so while using the failure probabilities of the good event \(G_t\), we get:

\[
\text{Reg}(T) = \sum_{t=1}^{T} \mathbb{E}[\Delta_{a_t}] \leq \sum_{t=1}^{T} \mathbb{E}[\Delta_{a_t}, \mathbb{1}\{G_t\}] + \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}\{\overline{G_t}\}, \mathbb{1}\{G_t\}]
\]

\[
= \sum_{a \notin \mathcal{A}_a} \Delta_a \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{G_t, a_t = a\}\right] + \Delta_{\text{max}} \sum_{t=1}^{T} \mathbb{P}r\{\overline{G_t}\} \leq \Delta_{\text{max}} C(T)
\]

\[
\leq \sum_{a \notin \mathcal{A}_a} \Delta_a \left(\bar{N}_a^g(T, \mu_a) + L_{\text{e}}(T, \Delta_a) + 1\right) + \Delta_{\text{max}} C(T)
\]

\[
\leq \sum_{a \notin \mathcal{A}_a} \Delta_a \left(\bar{N}_a^g(T, \mu_a) + L_{\text{e}}(T, \Delta_a)\right) + (K + C(T))\Delta_{\text{max}}
\]

Notice that substituting \(\mu_a\) leads to the desired bound, whether there is a unique or multiple optimal arms.

We similarly use Equation (17) to bound the expected number of queries. We still choose the same values of \(\mu_a\) for suboptimal arms, but for optimal arms, we let \(\mu_a = 0 = \Delta_a\) for all \(a \in \mathcal{A}_a\) when there are multiple optimal arms and \(\mu_a = \Delta_{\min}/2\) for a unique optimal arm. Then, as in the regret bound, we get

\[
\mathbb{E}[B^q(T)] = \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}\{q_t = 1\}] \leq \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}\{G_t, q_t = 1\}] + \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}\{\overline{G_t}\}, \mathbb{1}\{G_t\}]
\]

\[
= \sum_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{G_t, a_t = a, q_t = 1\}\right] + \sum_{t=1}^{T} \mathbb{P}r\{\overline{G_t}\} \leq \Delta_{\text{max}} C(T)
\]

\[
\leq \sum_{a \in [K]} \left(\bar{N}_a^g(T, \mu_a) + 1\right) + C(T)
\]

\[
= \sum_{a \in \mathcal{A}_a} \bar{N}_a^g(T, \mu_a) + \sum_{a \notin \mathcal{A}_a} \bar{N}_a^g(T, \mu_a) + K + C(T)
\]

and one can easily verify that substituting \(\mu_a\) leads to both desired bounds. \(\square\)
Remark 1. Notice that the bounds of (17) and (18) hold even if \( \epsilon(t) \) is \( F_t \) predictable, e.g., if the sequence \( \{\epsilon(t)\}_{t=1}^T \) is chosen by an adaptive adversary. Specifically, when this is the case, all bounds remain the same, except an expectation that should be taken on \( \bar{N}_g(a; T, \mu) \) and \( L_n(T, \Delta) \).

B.2 Problem-Independent Upper Bound

In this section, we generalize Proposition 4 to the setting presented in Appendix B.

Proposition 6. Under the notations of Theorem 3, let \( T \geq 1 \) and assume that \( \epsilon(t) \geq 0 \) for all \( t \in [T] \) and that Algorithm 1 is applied with regular confidence intervals. Also, assume that there exist a function \( M(t, \Delta_0) \) such that for all \( \Delta_0 > 0 \) and all arms \( a \in [K] \) with \( \Delta_a > \Delta_0 \), it holds that \( \Delta_a \bar{N}_g(a; T, \Delta_0) \leq M(T, \Delta_0) \). Then,

\[
\text{Reg}(T) \leq \inf_{\Delta_0 > 0} \left\{ 2KM \left( T, \frac{\Delta_0}{2} \right) + \Delta_0 T \right\} + \sum_{t=1}^{T} \epsilon(t) + (K + C(T))\Delta_{\text{max}} .
\]

See that for Hoeffding-based confidence intervals, we have that \( \bar{N}_g(a; T, \mu) \leq \frac{6 \ln T}{\Delta_a^2} \) (by Lemma 7), so for any \( \Delta_a > \Delta_0 \), it holds that \( \Delta_a \bar{N}_g(a; T, \Delta_0) = \frac{6 \ln T}{\Delta_a^2} \leq \frac{6 \ln T}{\Delta_0^2} \triangleq M(T, \Delta_0) \). Then, the infimum in the bound is achieved for \( \Delta_0 = \sqrt{\frac{24K \ln T}{\Delta_a}} \), which leads to the bound in the main paper (Proposition 4). Alternatively, when using Bernstein-type confidence bounds, we can bound the variance of arms by \( V_a \leq \frac{1}{4} \) and obtain a similar bound (by Lemma 8):

\[
\Delta_a \bar{N}_g(a; T, \Delta_0) \leq \frac{24V_a \ln T}{\Delta_a} + 52 \ln T + \Delta_a \leq \frac{6 \ln T}{\Delta_0} + 52 \ln T + \Delta_{\text{max}} \triangleq M(T, \Delta_0) .
\]

On the other hand, if the variance of all suboptimal arms is upper bounded by \( V_a \leq V \), this can be used to achieve an improved variance-dependent regret bound.

Proof. As in the problem-dependent bound of Theorem 3 we decompose the regret by:

\[
\text{Reg}(T) = \sum_{t=1}^{T} \mathbb{E}[\Delta_a] \leq \sum_{t=1}^{T} \mathbb{E}[\Delta_a, I\{G_t\}] + \Delta_{\text{max}} \sum_{t=1}^{T} \mathbb{E}[I\{G_t\}] \\
\leq \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E} \left[ \sum_{t=1}^{T} I\{G_t, a_t = a\} \right] + \Delta_{\text{max}} C(T) .
\]

Next, for any \( \Delta_0 > 0 \), we can bound the regret by

\[
\text{Reg}(T) \leq \sum_{a: \Delta_a > \Delta_0} \Delta_a \mathbb{E} \left[ \sum_{t=1}^{T} I\{G_t, a_t = a\} \right] + \sum_{a: \Delta_a \leq \Delta_0} \Delta_a \mathbb{E} \left[ \sum_{t=1}^{T} I\{G_t, a_t = a\} \right] \\
+ \Delta_{\text{max}} C(T) .
\]

To bound term (i) we further decompose it to

\[
(i) = \mathbb{E} \left[ \sum_{a: \Delta_a > \Delta_0} \Delta_a \sum_{t=1}^{T} I\{G_t, a_t = a, q_t = 1\} \right] + \mathbb{E} \left[ \sum_{a: \Delta_a > \Delta_0} \Delta_a \sum_{t=1}^{T} I\{G_t, a_t = a, q_t = 0\} \right] \\
\leq \mathbb{E} \left[ \sum_{a: \Delta_a > \Delta_0} \Delta_a \sum_{t=1}^{T} I\{G_t, a_t = a, q_t = 1\} \right] + \mathbb{E} \left[ \sum_{t=1}^{T} \Delta_a I\{G_t, q_t = 0\} \right] \\
+ \mathbb{E} \left[ \sum_{t=1}^{T} \Delta_a I\{G_t, q_t = 0\} \right] \\
+ \mathbb{E} \left[ \sum_{t=1}^{T} \Delta_a I\{G_t, q_t = 0\} \right] .
\]
For term (\(*\)), if there are multiple optimal arms, then by Equation (17) with \(\Delta = \Delta_a\),

\[
\sum_{a: \Delta_a > \Delta_0} \Delta_a \sum_{t=1}^T 1\{G_t, a_t = a, q_t = 1\} \leq \sum_{a: \Delta_a > \Delta_0} \Delta_a (\hat{N}_a^g(T, \Delta_a) + 1) \leq KM\left(T, \frac{\Delta_0}{2}\right) + K\Delta_{\text{max}},
\]

where the last inequality is since the gaps of all arms in the summation are larger than \(\Delta_0\), and specifically larger than \(\Delta_0/2\), and by bounding the number of arms with gaps larger than \(\Delta_0\) by \(K\). Alternatively, if there is a unique optimal arm, we can bound using Equation (17) with \(\Delta = \Delta_a/2\):

\[
\sum_{a: \Delta_a > \Delta_0} \Delta_a \sum_{t=1}^T 1\{G_t, a_t = a, q_t = 1\} \leq \sum_{a: \Delta_a > \Delta_0} \Delta_a \left(\hat{N}_a^g \left(T, \frac{\Delta_a}{2}\right) + 1\right)
\]

\[
\leq 2 \sum_{a: \Delta_a > \Delta_0} \frac{\Delta_a}{2} \hat{N}_a^g \left(T, \frac{\Delta_a}{2}\right) + K\Delta_{\text{max}}
\]

\[
\leq 2KM\left(T, \frac{\Delta_0}{2}\right) + K\Delta_{\text{max}},
\]

where the last inequality is again by the definition of \(M\left(T, \frac{\Delta_0}{2}\right)\). Combining both cases, we get

\[
\sum_{a: \Delta_a > \Delta_0} \Delta_a \sum_{t=1}^T 1\{G_t, a_t = a, q_t = 1\} \leq 2KM\left(T, \frac{\Delta_0}{2}\right) + K\Delta_{\text{max}}.
\]

The remaining summation (**) can be bounded using the fact that if \(q_t = 0\), then \(\Delta_a \leq \epsilon(t)\) (by Lemma [5]), and thus

\[
(**) = \sum_{t=1}^T \Delta_a 1\{G_t, q_t = 0\} \leq \sum_{t=1}^T \Delta_a \sum_{t=1}^T 1\{\Delta_a \leq \epsilon(t)\} \leq \sum_{t=1}^T \epsilon(t) 1\{\Delta_a \leq \epsilon(t)\} \leq \sum_{t=1}^T \epsilon(t) ,
\]

which lead to the bound of

\[
(i) \leq 2KM\left(T, \frac{\Delta_0}{2}\right) + \sum_{t=1}^T \epsilon(t) + K\Delta_{\text{max}}.
\]

Finally, we bound term (ii) as follows:

\[
\sum_{a: \Delta_a \in (0, \Delta_0]} \Delta_a \mathbb{E} \left[ \sum_{t=1}^T 1\{G_t, a_t = a\} \right] \leq \Delta_0 \mathbb{E} \left[ \sum_{a=1}^K \sum_{t=1}^T 1\{G_t, a_t = a\} \right] = \Delta_0 \mathbb{E} \left[ \sum_{t=1}^T 1\{G_t\} \right] \leq \Delta_0 T.
\]

Substituting (i) and (ii) into Equation (19) and taking the infimum over all possible choices of \(\Delta_0 > 0\) leads to the desired bound. \qed
B.3 Regularity Proofs

B.3.1 Hoeffding-Based Confidence Intervals

Hoeffding-based confidence intervals are probably the most commonly used confidence intervals for bounded (and subgaussian) rewards. Specifically, for rewards bounded in $[0, 1]$, they are defined by

$$UCB_t(a) = \hat{\mu}_{t-1}(a) + \sqrt{\frac{3 \ln t}{2n_t^q(a)}}$$
and
$$LCB_t(a) = \hat{\mu}_{t-1}(a) - \sqrt{\frac{3 \ln t}{2n_t^q(a)}},$$

and if $n_t^q(a) = 0$, we define $UCB_t(a) = +\infty$ and $LCB_t(a) = -\infty$. In the following, we prove that such confidence intervals are regular:

**Lemma 7.** Assume that the rewards are bounded in $[0, 1]$. Then, for any $T \geq 1$, Hoeffding-based confidence intervals are regular w.r.t. the events $G_t = \{\forall a \in [K] : \mu_a \in [LCB_t(a), UCB_t(a)]\}$ and the functions

$$C(t) = 2K, \quad N^a(t, \mu) = N^{a*}(t, \mu) = \min \left\{ \frac{6 \ln t}{\mu^2}, t - 1 \right\},$$

Specifically, $	ilde{N}^a_0(T, 0) = \tilde{N}^{a*}(T)$. Moreover, for any $\mu > 0$, it holds that $\tilde{N}^a_0(T, \mu) \leq \frac{6 \ln T}{\mu^2}$. Finally, if $B(t)$ is a positive nondecreasing sequence and $\epsilon(t) = \sqrt{\frac{6 \ln T}{B(t)}}$, or, alternatively, $\epsilon(t)$ is a nonnegative nonincreasing sequence, then $\tilde{N}^a_0(T, \mu) \leq \min \left\{ \frac{6 \ln T}{\max(\mu^2, \epsilon(t)^2)}, T - 1 \right\}$.

**Proof.** Notice that $CI_t(a) = \sqrt{\frac{6 \ln t}{n_t^q(a)}}$. We now check the conditions by their order:

1. The condition that $CI_t(a) \geq 0$ trivially holds for any $n_t^q(a) \geq 1$ and holds by definition when $n_t^q(a) = 0$ (and $CI_t(a) = \infty$).

2. By definition, for any $t \geq 1$, if $G_t$ holds, then $\mu_a \in [LCB_t(a), UCB_t(a)]$ for all $a \in [K]$, as required.

3. We now bound $\Pr \{ G_t \}$. For any $a \in [K]$, let $X_a(1), \ldots X_a(T)$ be i.i.d. random variables of the same distribution as arm $a$, and we let the observed reward at the $n_t^q(a)$th time the arm was played be $R_t = X_a(n_t^q(a))$. Then, we can write:

$$\Pr \{ G_t \} = \Pr \{ \exists a \in [K] : \mu_a \notin [LCB_t(a), UCB_t(a)] \}$$

$$= \Pr \left\{ \exists a \in [K] : |\hat{\mu}_{t-1}(a) - \mu_a| > \sqrt{\frac{3 \ln t}{2n_t^q(a)}} \right\}$$

$$\leq \sum_{a=1}^{K} \Pr \left\{ |\hat{\mu}_{t-1}(a) - \mu_a| > \sqrt{\frac{3 \ln t}{2n_t^q(a)}} \right\}$$

$$\leq \sum_{a=1}^{K} \sum_{n=1}^{t-1} \Pr \left\{ |\hat{\mu}_{t-1}(a) - \mu_a| > \sqrt{\frac{3 \ln t}{2n_t^q(a)}}, n_t^q(a) = n \right\}$$

$$\leq \sum_{a=1}^{K} \sum_{n=1}^{t-1} \Pr \left\{ \left| \frac{1}{n} \sum_{k=1}^{n} X_a(k) - \mu_a \right| > \sqrt{\frac{3 \ln t}{2n}}, n_t^q(a) = n \right\}$$

$$\leq \sum_{a=1}^{K} \sum_{n=1}^{t-1} \Pr \left\{ \left| \frac{1}{n} \sum_{k=1}^{n} X_a(k) - \mu_a \right| > \sqrt{\frac{3 \ln t}{2n}} \right\}$$

$$\leq \frac{K}{t^2}. $$
We now prove the claims at the end of the lemma. First, for any nonnegative sequence \( a \in [K] \) and \( n_{t-1}^a(a) \in \{0, \ldots, t-1\} \), respectively. Moreover, notice that the relevant event cannot occur when \( n_{t-1}^a(a) = 0 \), so we only treat values larger than 1. In (3) we substituted \( n_{t-1}^a(a) = n \) and wrote the empirical means using \( X_n(t) \). Finally, (4) relies on Hoeffding’s inequality for i.i.d. bounded random variables supported by \([0, 1]\). Summing over all time indices leads to the desired result:

\[
\sum_{t=1}^{T} \Pr \{ \bar{G}_t \} \leq \sum_{t=1}^{T} \frac{K}{t^2} \leq \sum_{t=1}^{\infty} \frac{K}{t^2} \leq 2K = C(T) .
\]

4. If \( N^a(t, \mu) = t - 1 \), then the event that \( N^a(t, \mu) < n_{t-1}^a(a) \leq t - 1 \) can never happen, so assume w.l.o.g. that \( N^a(t, \mu) < t - 1 \) and \( N^a(t, \mu) = \frac{6\ln t}{\mu^2} \). In particular, \( \mu > 0 \) and \( t > 1 \), and it can be easily verified that whether the good event holds or not, if \( n_{t-1}^a(a) > N^a(t, \mu) = \frac{6\ln t}{\mu^2} \), then

\[
CI_a(a) = \sqrt{\frac{6\ln t}{n_{t-1}^a(a)}} < \sqrt{\frac{6\ln t}{N^a(t, \mu)}} = \mu .
\]

5. The proof follows exactly as the previous part, as its result did not depend on the good event.

We now prove the claims at the end of the lemma. First, for any nonnegative sequence \( \{\epsilon(t)\}_{t \geq 1} \), it holds that

\[
\bar{N}^a(t, \mu) = \max_{t \in [T]} N^a(t, \max \{ \mu, \epsilon(t) \})
\]

\[
= \max_{t \in [T]} \left\{ \min \left\{ \frac{6\ln t}{\max \{ \mu^2, \epsilon^2(t) \}}, t - 1 \right\} \right\}
\]

\[
\leq \max_{t \in [T]} \left\{ \min \left\{ \frac{6\ln t}{\mu^2}, t - 1 \right\} \right\}
\]

\[
\leq \min \left\{ \frac{6\ln T}{\mu^2}, T - 1 \right\}
\]

where the last inequality is since the logarithmic and constant functions are nondecreasing in \( t \), and the minimum of nondecreasing functions is nondecreasing (Lemma 9). Finally, we write

\[
N^a(t, \max \{ \mu, \epsilon(t) \}) = \min \left\{ \frac{6\ln t}{\mu^2}, \frac{6\ln t}{\epsilon^2(t)}, t - 1 \right\} .
\]

Moreover, if either \( \epsilon(t) = \sqrt{\frac{6\ln t}{2B(t)}} \), for nondecreasing positive \( B(t) \), or \( \epsilon(t) \geq 0 \) is nonincreasing, then \( \frac{6\ln t}{\epsilon^2(t)} \) is nondecreasing in \( t \). Thus, \( N^a(t, \max \{ \mu, \epsilon(t) \}) \) is a minimum of nondecreasing functions and is nondecreasing by itself (by Lemma 9). In turn, this implies that

\[
\bar{N}^a(T, \mu) = \max_{t \in [T]} N^a(t, \max \{ \mu, \epsilon(t) \}) = N^a(T, \max \{ \mu, \epsilon(T) \})
\]

\[
= \min \left\{ \frac{6\ln T}{\max \{ \mu^2, \epsilon^2(T) \}}, T - 1 \right\} .
\]
**B.3.2 Bernstein-Based Confidence Intervals**

Bernstein-based confidence intervals are confidence intervals that depend on the variance on the reward distributions. Specifically, we apply Empirical-Bernstein bounds [Maurer and Pontil 2009], which depend on the empirical variance of arms and obviate the need to know the true variances of arms. Formally, if all arm distributions are bounded in $[0, 1]$ and of variances $\{V_a\}_{a \in [K]}$, we denote the unbiased empirical estimator of the variance by

$$
\hat{V}_t(a) = \frac{n^a_t}{n^g_t} \left( \frac{1}{n^g_t} \sum_{i=1}^{t} R_i^a \mathbb{1}\{a_i = a\} - (\hat{\mu}_t(a))^2 \right),
$$

where we define $\hat{V}_t(a) = 0$ if $n^g_t(a) < 2$. Then, Bernstein-based confidence intervals are defined by

$$
UCB_t(a) = \hat{\mu}_{t-1}(a) + \sqrt{\frac{6\hat{V}_t(a) \ln t}{n^g_{t-1}(a)}} + \frac{7 \ln t}{n^g_{t-1}(a) - 1},
$$

$$
LCB_t(a) = \hat{\mu}_{t-1}(a) - \sqrt{\frac{6\hat{V}_t(a) \ln t}{n^g_{t-1}(a)}} - \frac{7 \ln t}{n^g_{t-1}(a) - 1},
$$

where if $n^g_{t-1}(a) \leq 1$, we define $UCB_t(a) = +\infty$ and $LCB_t(a) = -\infty$ (and in general, we let $\frac{\ln t}{n^g(a) - 1} = +\infty$ for $n^g_t(a) \leq 1$). We now prove that these confidence intervals are regular:

**Lemma 8.** Assume that the rewards are bounded in $[0, 1]$ and let $T \geq 1$. Then, Bernstein-based confidence interval are regular w.r.t. the events

$$
\mathbb{G}_t = \left\{ \forall a \in [K] : \mu_a \in [LCB_t(a), UCB_t(a)], \left| \sqrt{\hat{V}_t(a)} - \sqrt{V_a} \right| \leq \sqrt{\frac{6 \ln t}{n^g_{t-1}(a) - 1}} \right\}
$$

and the functions

$$
C(t) = 12K, \quad N^g_a(t, \mu) = \min \left\{ \frac{24V_a \ln t}{\mu^2} + \frac{52 \ln t}{\mu} + 1, t-1 \right\}.
$$

For $N^{as}(t, \mu)$, we allow two different options:

$$
N^{as}_1(t, \mu) = \min \left\{ \frac{6 \ln t}{\mu^2} + \frac{28 \ln t}{\mu} + 1, t-1 \right\}, \quad \text{or}
$$

$$
N^{as}_2(t, \mu) = \min \left\{ \left( \frac{3 \ln t}{2\mu^2} + \sqrt{\frac{14 \ln t}{\mu^2}} \right)^2 + 1, t-1 \right\}
$$

Specifically, we have that $\bar{N}^g_a(T, \mu) \leq N^g_a(T, \mu)$. Finally, we suggest two choices for $\epsilon(t)$:

- For nonincreasing sequence $\epsilon(t) \geq 0$, we have that $\bar{N}^g_a(T, \mu) = N^g_a(T, \mu)$ and $\bar{N}^{as}_i(T) = N^{as}_i(T, \mu)$ for $i = 1, 2$.

- If $B(t)$ is a nondecreasing sequence such that $B(1) > 1$ and $\epsilon(t) = \sqrt{\frac{6 \ln t}{B(t)} + 14 \ln t}$, we work with $N^{as}_2(t, \epsilon(t)) = \min \{B(t), t-1\}$, and thus, $\bar{N}^{as}_i(T) = \min \{B(T), T-1\}$. Moreover, for this choice, we have that

$$
\bar{N}^g_a(T, \mu) \leq \min \left\{ N^g_a(T, \mu), 4V_a(B(T)-1) + \min \left\{ 22 \sqrt{(B(T)-1) \ln T}, 4(B(T)-1) \right\} + 1 \right\}.
$$

**Proof.** First note that for any $a \in [K]$ and $t \geq 1$, the width of the confidence interval is

$$
CI_t(a) = 2 \sqrt{\frac{6\hat{V}_t(a) \ln t}{n^g_{t-1}(a)}} + 14 \frac{\ln t}{n^g_{t-1}(a) - 1}.
$$

We now verify each of the regularity requirements:
1. The condition that $CI_t(a) \geq 0$ trivially holds for any $t \geq 1$ when $n_{t-1}^q(a) \geq 2$, as $\hat{V}_{t-1}(a) \geq 0$ and $\ln T \geq 0$, and hold by definition when $n_{t-1}^q(a) \leq 1$, as then, $CI_t(a) = +\infty$.

2. By definition, under $G_t$, $\mu_a \in [LCB_t(a), UCB_t(a)]$ for any $t \geq 1$ and $a \in [K]$.

3. We now bound $\Pr\{G_t\}$. For any $a \in [K]$, let $X_a(1), \ldots, X_a(T)$ be i.i.d. random variables of the same distribution as arm $a$, and we let the observed reward at the $n_t^q(a)$th time the arm was played be $R_t = X_a(n_t^q(a))$. We also denote the unbiased empirical estimator of the variance based on the first $n$ samples by

$$
\hat{V}_n(a) = \frac{n}{n-1} \left( \frac{1}{n} \sum_{k=1}^{n} X_a^2(k) - \left( \frac{1}{n} \sum_{k=1}^{n} X_a(k) \right)^2 \right),
$$

which equals $\hat{V}_{t-1}(a)$ when $n_{t-1}^q(a) = n$. Then, we can write:

$$
\Pr\{G_t\} = \Pr\left\{ \exists a \in [K] : \mu_a \notin [LCB_t(a), UCB_t(a)], \quad \text{or} \right. $$

$$
\left| \sqrt{\hat{V}_{t-1}(a)} - \sqrt{V_a} \right| > \sqrt{\frac{6 \ln t}{n_{t-1}^q(a) - 1}}
$$

$$
= \Pr\left\{ \exists a \in [K] : |\hat{\mu}_{t-1}(a) - \mu_a| > \sqrt{\frac{6 \hat{V}_{t-1}(a) \ln t}{n_{t-1}^q(a)}} + \frac{7 \ln t}{n_{t-1}^q(a) - 1}, \quad \text{or} \right. $$

$$
\left| \sqrt{\hat{V}_{t-1}(a)} - \sqrt{V_a} \right| > \sqrt{\frac{6 \ln t}{n_{t-1}^q(a) - 1}}
$$

$$(1) \quad \sum_{a=1}^{K} \sum_{n=2}^{t-1} \Pr\left\{ |\hat{\mu}_{t-1}(a) - \mu_a| > \sqrt{\frac{6 \hat{V}_{t-1}(a) \ln t}{n_{t-1}^q(a)}} + \frac{7 \ln t}{n_{t-1}^q(a) - 1}, n_{t-1}^q(a) = n \right\}

+ \sum_{a=1}^{K} \sum_{n=2}^{t-1} \Pr\left\{ \left| \sqrt{\hat{V}_{t-1}(a)} - \sqrt{V_a} \right| > \sqrt{\frac{6 \ln t}{n_{t-1}^q(a) - 1}}, n_{t-1}^q(a) = n \right\}

$$(2) \quad \sum_{a=1}^{K} \sum_{n=2}^{t-1} \Pr\left\{ \left| \frac{1}{n} \sum_{k=1}^{n} X_a(k) - \mu_a \right| > \sqrt{\frac{6 \hat{V}_n(a) \ln t}{n}} + \frac{7 \ln t}{n-1}, n_{t-1}^q(a) = n \right\}

+ \sum_{a=1}^{K} \sum_{n=2}^{t-1} \Pr\left\{ \left| \sqrt{\hat{V}_n(a)} - \sqrt{V_a} \right| > \sqrt{\frac{6 \ln t}{n-1}}, n_{t-1}^q(a) = n \right\}

$$

$$(3) \quad \sum_{a=1}^{K} \sum_{n=1}^{t-1} \frac{4}{\sqrt{n}} + \sum_{a=1}^{K} \sum_{n=1}^{t-1} \frac{2}{\sqrt{n}} \leq \frac{6K}{t^2} \quad .
$$

Relations (1) is due to the union bound over the two events, actions $a \in [K]$ and $n_{t-1}^q(a) \in \{0, \ldots, t-1\}$, respectively (as in the more detailed proof of Lemma [7]). Specifically, notice that the relevant events cannot occur when $n_{t-1}^q(a) \leq 1$, so we only treat values larger than 2. In (2) we substituted $n_{t-1}^q(a) = n$ and wrote the empirical means and variance using $X_a(n)$ and $\hat{V}_n(n)$. Finally, (3) relies on Theorems 4.
5. As in the proof of 4. If

\[ \sum_{i=1}^{T} \text{Pr}\{\mathcal{G}_i\} \leq \sum_{i=1}^{T} \frac{6K}{t^2} \leq \sum_{i=1}^{\infty} \frac{6K}{t^2} \leq 12K = C(T). \]

4. If \( N_o^g(t, \mu) = t - 1 \), then the event that \( N_o^g(t, \mu) < n_t^q(a) \leq t - 1 \) can never happen, so assume w.l.o.g. that \( N_o^g(t, \mu) < t - 1 \) and, in particular, \( t > 1, \mu > 0 \) and \( N_o^g(t, \mu) = \frac{24V_o \ln t}{\mu^2} + \frac{52 \ln t}{\mu} + 1 \). Notice that in this case, \( N_o^g(t) > 1 \), and for \( n_t^q(a) > N_o^g(t) \), all bounds are finite and all denominators are positive.

Also note that under \( \mathcal{G}_t \), we can bound \( \sqrt{V_{t-1}(a)} \leq \sqrt{V_a} + \sqrt{\frac{6\ln t}{n_t^q(a)-1}} \). Then, substituting to \( CI_t(a) \), for any \( t > 1 \) and \( a \in [K] \),

\[
CI_t(a) = 2\sqrt{\frac{6V_{t-1}(a) \ln t}{n_t^q(a)} + 14 \frac{\ln t}{n_t^q(a)-1}}.
\]

Finally, by Lemma 10 (presented below), see that if \( n_t^q(a) > \frac{24V_o \ln t}{\mu^2} + \frac{52 \ln t}{\mu} + 1 = N_o^g(t, \mu) \), then \( CI_t(a) < \mu \), as required.

5. As in the proof of \( N_o^g(t, \mu) \), we focus on the case when \( N_i^{g*}(t, \mu) < t - 1 \), which implies that \( t > 1 \) and \( \mu > 0 \) (for both choices of \( i \in \{1, 2\} \)). Thus, for \( n_t^q(a) > N_i^{g*}(t, \mu) > 1 \), all bounds are finite and all denominators are positive. Also, for rewards bounded in \([0, 1]\), we can bound \( \hat{V}_t(a) \) by

\[
\hat{V}_t(a) = \frac{n_t^q(a)}{n_t^q(a) - 1} \left( \frac{1}{n_t^q(a)} \sum_{s=1}^{t} R_t^s \mathbf{1}\{a_t = a\} - (\hat{\mu}_t(a))^2 \right)
\leq \frac{n_t^q(a)}{n_t^q(a) - 1} \left( \frac{1}{n_t^q(a)} \sum_{s=1}^{t} R_t \mathbf{1}\{a_t = a\} - (\hat{\mu}_t(a))^2 \right)
= \frac{n_t^q(a)}{n_t^q(a) - 1} (\hat{\mu}_t(a) - (\hat{\mu}_t(a))^2)
\leq \frac{1}{4} \frac{n_t^q(a)}{n_t^q(a) - 1},
\]

where the last inequality is since the function \( x - x^2 \leq \frac{1}{4} \) for \( x \in [0, 1] \). In turn, for all \( n_t^q(a) > 1 \),

\[
CI_t(a) = 2\sqrt{\frac{6V_{t-1}(a) \ln t}{n_t^q(a)} + 14 \frac{\ln t}{n_t^q(a)-1}}
\leq \sqrt{\frac{6\ln t}{n_t^q(a) - 1} + 14 \frac{\ln t}{n_t^q(a)-1}}.
\]

By elementary algebra, if \( n_t^q(a) = \left( \sqrt{\frac{3\ln t}{2\mu^2}} + \sqrt{\frac{3\ln t}{2\mu^2} + \frac{14\ln t}{\mu}} \right)^2 + 1 = N_o^{g*}(t, \mu) \), then the above bound exactly equals to \( \mu \). Moreover, the bound strictly decreases in \( n_t^q(a) \); therefore, for any

\[ 33 \]
\( n_{t-1}^a(a) > N_{i}^{a\alpha}(t, \mu) \), we get that \( CI_i(a) < \mu \), as required of \( N_{i}^{a\alpha} \). Alternatively, by Lemma \[10\] if \( n_{t-1}^a(a) > \frac{6\ln t}{\mu^2} + \frac{28\ln t}{\mu} + 1 = N_{i}^{a\alpha}(t, \mu) \), then \( CI_i(a) < \mu \), as desired from \( N_{i}^{a\alpha} \).

We now prove the additional results stated at the end of the lemma. First, notice that

\[
\bar{g}(N_a(t, \max\{\mu, \epsilon(t)\})) = \min \left\{ \frac{24V_a \ln t}{\max\{\mu^2, \epsilon(t)\}} + \frac{52\ln t}{\max\{\mu, \epsilon(t)\}} + 1, t-1 \right\}
\]

(20)

= \min \left\{ \frac{24V_a \ln t}{\mu^2} + \frac{52\ln t}{\mu} + 1, \frac{24V_a \ln t}{\epsilon(t)} + \frac{52\ln t}{\epsilon(t)} + 1, t-1 \right\}

\leq \min \left\{ \frac{24V_a \ln t}{\mu^2} + \frac{52\ln t}{\mu} + 1, t-1 \right\}

= N_a^g(t, \mu) .

Moreover, \( N_a^g(t, \mu) \) is nondecreasing in \( t \), and thus,

\[
\bar{g}(N_a(T, \mu)) = \max_{t \in [T]} N_a^g(t, \max\{\mu, \epsilon(t)\}) \leq \max_{t \in [T]} N_a^g(t, \mu) \leq N_a^g(T, \mu) .
\]

Next, if \( \epsilon(t) \) is nonincreasing, then Equation (20) expresses \( N_a^g(t, \max\{\mu, \epsilon(t)\}) \) as a minimum on nondecreasing functions, which is by itself nondecreasing (see Lemma \[9\]). Thus, in this case, it holds that

\[
\bar{g}(N_a(T, \mu)) = \max_{t \in [T]} N_a^g(t, \max\{\mu, \epsilon(t)\}) = N_a^g(T, \max\{\mu, \epsilon(T)\}) .
\]

One can similarly verify that when \( \epsilon(t) \) is nonincreasing, \( N_a^{a\alpha}(t, \epsilon(t)) \) are minima of nondecreasing functions, and following similar lines implies that \( \bar{g}(N_i^{a\alpha}(T, \epsilon(T))) \leq N_i^{a\alpha}(T, \epsilon(T)) \) for \( i = 1, 2 \).

Finally, assume that \( B(t) > 1 \) is nondecreasing and let \( \epsilon(t) = \sqrt{\frac{6\ln t}{B(t)-1}} + 14\ln t/(B(t)-1) \). Then, a direct calculation results with \( \bar{g}(N_a^{a\alpha}(t, \epsilon(t))) = \min\{B(t), t-1\} \), which is also nondecreasing, a property that once again implies that \( \bar{g}(N_a^{a\alpha}(T)) = \min\{B(t), t-1\} \). Moreover, returning to (20), we can write

\[
N_a^g(t, \max\{\mu, \epsilon(t)\}) = \min \left\{ \frac{24V_a \ln t}{\epsilon(t)} + \frac{52\ln t}{\epsilon(t)} + 1, \frac{24V_a \ln t}{\mu^2} + \frac{52\ln t}{\mu} + 1, t-1 \right\}
\]

We now substitute \( \epsilon(t) \) and bound the second term by

\[
\frac{24V_a \ln t}{\epsilon(t)} + \frac{52\ln t}{\epsilon(t)} \leq \frac{24V_a \ln t}{(\sqrt{6\ln t/(B(t)-1)}) + 14\ln t/(B(t)-1)}\]

\[
\leq 4V_a(B(t)-1) + \text{min}\left\{ 22\sqrt{(B(t)-1)\ln t}, 4(B(t)-1) \right\} .
\]

Importantly, as \( B(t) \) is nondecreasing, this bound is nondecreasing. In turn, substituting back to \( N_a^g(t, \max\{\mu, \epsilon(t)\}) \) results with a nondecreasing upper bound, whose maximum is achieved at \( t = T \). This leads to the desired bound on \( N_a^g(T, \mu) \). \( \square \)

### B.3.3 Auxiliary Lemmas

We now present two extremely simple results that we repeatedly used and are proven for completeness:

**Lemma 9.** Let \( f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R} \) be nondecreasing functions. Then, \( f(x) = \min_{j \in [n]} f_j(x) \) is also nondecreasing in \( x \).

**Proof.** Let \( x_1, x_2 \in \mathbb{R} \) such that \( x_1 > x_2 \) and assume w.l.o.g. that \( f(x_1) = f_k(x_1) \) for some \( k \in [n] \). Then,

\[
f(x_1) = f_k(x_1) \geq f_k(x_2) \geq \min_{j \in [n]} f_j(x_2) = f(x_2) .
\]

\( \square \)
Lemma 10. If $c_1, c_2 \geq 0$ and $\mu > 0$, then for any $n > n_0 \triangleq \frac{c_1^2}{\mu^2} + \frac{2c_2}{\mu}$, it holds that

$$\frac{c_1}{\sqrt{n}} + \frac{c_2}{n} < \mu.$$ 

Proof. Notice that the l.h.s. is strictly increasing in $n$. Therefore, it is sufficient to prove that for $n = n_0$, the l.h.s. is (weakly) smaller than $\mu$:

$$\frac{c_1}{\sqrt{n_0}} + \frac{c_2}{n_0} = \frac{c_1}{\sqrt{n_0}} + \frac{c_2}{n_0} = \frac{c_1}{\sqrt{n_0}} + \frac{c_2}{n_0} + \frac{c_1}{\sqrt{n_0}} + \frac{c_2}{n_0} = \mu \cdot \frac{c_1}{\sqrt{n_0}} + \frac{2c_2}{\mu} + \frac{c_2}{\mu} = \mu \cdot \frac{c_1}{\sqrt{n_0}} + \frac{2c_2}{\mu} + \frac{c_2}{\mu} = \mu \cdot \frac{c_1}{\sqrt{n_0}} + \frac{2c_2}{\mu} + \frac{c_2}{\mu} = \mu$$

where $(\ast)$ is by the inequality $\sqrt{a(a + b)} \leq a + \frac{b}{2}$. \hfill $\Box$
C Experimental details

C.1 Baseline Algorithms

Algorithm 2 Budget-Feedback Aware Upper Confidence Bound (BuFAU)

1: Define: \( UCB_t(a) = \hat{\mu}_{t-1}(a) + \sqrt{\frac{3 \ln t}{2n^2_{t-1}(a)}} \) and \( LCB_t(a) = \hat{\mu}_{t-1}(a) - \sqrt{\frac{3 \ln t}{2n^2_{t-1}(a)}} \)
2: for \( t = 1, \ldots, K \) do
3: \hspace{1em} Play \( a_t = t \) and ask for feedback \( (q_t = 1) \)
4: \hspace{1em} Observe \( R_t \) and update \( n^2_t(a_t), \hat{\mu}_t(a_t) \)
5: end for
6: for \( t = K + 1, \ldots, T \) do
7: \hspace{1em} Observe \( \epsilon(t) \geq 0 \)
8: \hspace{1em} Set \( l_t \in \arg \max_a LCB_t(a) \) and \( u_t \in \arg \max_a UCB_t(a) \)
9: \hspace{1em} if \( \max_{a \neq l_t} UCB_t(a) \leq LCB_t(l_t) \) or \( UCB_t(u_t) - LCB_t(l_t) \leq \epsilon(t) \) then
10: \hspace{2em} Play \( a_t = l_t \) and do not ask for feedback \( (q_t = 0) \)
11: \hspace{1em} else
12: \hspace{2em} Play \( a_t = u_t \) and ask for feedback \( (q_t = 1) \)
13: \hspace{1em} Observe \( R_t \) and update \( n^2_t(a_t), \hat{\mu}_t(a_t) \)
14: end if
15: end for

Algorithm 3 Confidence-Budget Matching Upper Confidence Bound (CBM-UCB)

1: Define: \( UCB_t(a) = \hat{\mu}_{t-1}(a) + \sqrt{\frac{3 \ln t}{2n^2_{t-1}(a)}} \) and \( LCB_t(a) = \hat{\mu}_{t-1}(a) - \sqrt{\frac{3 \ln t}{2n^2_{t-1}(a)}} \)
2: for \( t = 1, \ldots, K \) do
3: \hspace{1em} Play \( a_t = t \) and ask for feedback \( (q_t = 1) \)
4: \hspace{1em} Observe \( R_t \) and update \( n^2_t(a_t), \hat{\mu}_t(a_t) \)
5: end for
6: for \( t = K + 1, \ldots, T \) do
7: \hspace{1em} Observe \( \epsilon(t) \geq 0 \)
8: \hspace{1em} Play \( a_t = u_t \in \arg \max_a UCB_t(a) \)
9: \hspace{1em} if \( CI_t(u_t) \leq \epsilon(t) \) then
10: \hspace{2em} Do not ask for feedback \( (q_t = 0) \)
11: \hspace{1em} else
12: \hspace{2em} Ask for feedback \( (q_t = 1) \), observe \( R_t \) and update \( n^2_t(a_t), \hat{\mu}_t(a_t) \)
13: end if
14: end for

Algorithm 4 Greedy Algorithm

1: Define: \( UCB_t(a) = \hat{\mu}_{t-1}(a) + \sqrt{\frac{3 \ln t}{2n^2_{t-1}(a)}} \), \( LCB_t(a) = \hat{\mu}_{t-1}(a) - \sqrt{\frac{3 \ln t}{2n^2_{t-1}(a)}} \) and \( B^g(K) = K \)
2: for \( t = 1, \ldots, K \) do
3: \hspace{1em} Play \( a_t = t \) and ask for feedback \( (q_t = 1) \)
4: \hspace{1em} Observe \( R_t \) and update \( n^2_t(a_t), \hat{\mu}_t(a_t) \)
5: end for
6: for \( t = K + 1, \ldots, T \) do
7: \hspace{1em} Observe \( \epsilon(t) \geq 0 \) and calculate effective budget \( B(t) = \frac{6K \ln t}{\epsilon^2(t)} + K \)
8: \hspace{1em} if \( B^g(t-1) > B(t) - 1 \) then
9: \hspace{2em} Play \( a_t \in \arg \max_a \hat{\mu}_{t-1}(a) \), do not ask for feedback \( (q_t = 0) \) and set \( B^g(t) = B^g(t-1) \)
10: \hspace{1em} else
11: \hspace{2em} Play \( a_t \in \arg \max_a UCB_t(a) \), ask for feedback \( (q_t = 1) \) and set \( B^g(t) = B^g(t-1) + 1 \)
12: \hspace{1em} Observe \( R_t \) and update \( n^2_t(a_t), \hat{\mu}_t(a_t) \)
13: end if
14: end for
C.2 Additional Experiments

C.2.1 Experiments in 5-Armed Random Problems

In this appendix, we present simulation results that demonstrate that the behavior presented in Section 4.1 still holds when arms are random. Specifically, we experiment on 5-armed problems with either one or two optimal arms, where the optimal mean is \( \mu^* = 0.5 \) and the suboptimal arms are of mean \( \mu_a = 0.25 \). All arms are Bernoulli-distributed. As in the main paper, we evaluate the algorithms for 100,000 time steps using \( \epsilon(t) = t^{-1/4} \). The simulation results in Figure 2 are averaged over 1,000 seeds, and the statistics at the last time-step are presented in Table 1 (including mean, standard deviation, 90\(^{th}\) quantile and maximum). The conclusions practically remain the same as in the main paper – all algorithms perform roughly the same when there are multiple optimal arms (with slight performance degradation and increased querying for the greedy algorithm). On the other hand, when there is a unique optimal arm, BuFALU requires much less feedback but suffers from a regret degradation by a factor of 3 (similar to the 4 factor of Theorem 2).

Figure 2: Evaluation of all algorithms on 5-armed problems over 1,000 seeds with \( \epsilon(t) = t^{-1/4} \).

For completeness, we also present simulation results for two additional profiles of \( \epsilon(t) \) – the first is when \( \epsilon(t) = 0 \) (see Figure 3 and statistics in Table 2) and the second when \( \epsilon(t) = \frac{1}{\ln t} \) (see Figure 4 and statistics in Table 3). When \( \epsilon(t) = 0 \), algorithms are allowed to ask as much feedback as they want. Then, when there are multiple optimal arms, all algorithms always ask for feedback and perform the same. On the other hand, when the optimal arm is unique, BuFALU asks for feedback on \( \sim \frac{1}{2} \) of the rounds and suffers a factor of 4 in its regret. Therefore, in the unlimited budget scenario, we return to the same behavior as in Section 4.1. The case of \( \epsilon(t) = \frac{1}{\ln t} \), which corresponds to a querying budget of \( O(K \ln^3 t) \), behaves very similar to the choice of \( \epsilon(t) = t^{-1/4} \). Interestingly, when the optimal arm is unique, the regret of BuFALU in both limited cases is lower than the regret when \( \epsilon(t) = 0 \). A possible explanation is that when the budget is unlimited, BuFALU aggressively explores to separate the optimal arm from all other arms. When the budget is limited, BuFALU combines exploration with exploitation of \( l_t \), which leads to a slightly lower regret.
Figure 3: Evaluation of all algorithms on 5-armed problems over 1,000 seeds with $\epsilon(t) = 0$.

Figure 4: Evaluation of all algorithms on 5-armed problems over 1,000 seeds with $\epsilon(t) = \frac{1}{\ln T}$.
Table 1: Additional statistics of the empirical evaluation in Figure 2. All statistics are measured after $T = 100,000$ time steps over 1,000 different seeds. All experiments use $\epsilon(t) = t^{-1/2}$. std is the standard deviation, 90% represents the 90th percentile and max is the maximum.

| Bandit Instance | Algorithm | Regret | Queries |
|-----------------|-----------|--------|---------|
|                 | mean      | std    | 90%     | max      | mean    | std    | 90%     | max      |
| 5-arms, unique optimal, values: $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$ | BuFALU    | 664.88 | 87.08   | 774.8    | 982.25  | 3471.38 | 437.39  | 4024.1   | 5023     |
|                 | BuFAU     | 222.94 | 23.03   | 253.25   | 304.5   | 2273.67 | 92.1    | 2285.8   | 23063    |
|                 | greedy    | 248.08 | 26.18   | 281.77   | 343.75  | 100000  | 0       | 100000   | 100000   |
|                 | CBM       | 222.97 | 23.06   | 253.25   | 304.5   | 2273.69 | 92.25   | 2285.8   | 23063    |
| 5-arms, multiple optimal, values: $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$ | BuFALU    | 167.64 | 20.17   | 194.52   | 230.75  | 38049.3 | 3944.21 | 42593.5  | 44135    |
|                 | BuFAU     | 164.62 | 19.75   | 190.02   | 230.25  | 39254.9 | 3292.47 | 43298.1  | 44383    |
|                 | greedy    | 178.79 | 21.72   | 207.78   | 256.75  | 100000  | 0       | 100000   | 100000   |
|                 | CBM       | 164.92 | 19.95   | 190.5    | 245     | 39217.6 | 3277.71 | 43289.4  | 44383    |

Table 2: Additional statistics of the empirical evaluation in Figure 3. All statistics are measured after $T = 100,000$ time steps over 1,000 different seeds. All experiments use $\epsilon(t) = 0$. std is the standard deviation, 90% represents the 90th percentile and max is the maximum.

| Bandit Instance | Algorithm | Regret | Queries |
|-----------------|-----------|--------|---------|
|                 | mean      | std    | 90%     | max      | mean    | std    | 90%     | max      |
| 5-arms, unique optimal, values: $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$ | BuFALU    | 1003.96| 131.23  | 1180.55  | 1457.25 | 3240.73 | 658.93  | 6130     | 7538     |
|                 | BuFAU     | 248.08 | 26.18   | 281.77   | 343.75  | 100000  | 0       | 100000   | 100000   |
|                 | greedy    | 248.08 | 26.18   | 281.77   | 343.75  | 100000  | 0       | 100000   | 100000   |
|                 | CBM       | 248.08 | 26.18   | 281.77   | 343.75  | 100000  | 0       | 100000   | 100000   |
| 5-arms, multiple optimal, values: $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$ | BuFALU    | 181.05 | 21.6    | 208.02   | 259.5   | 100000  | 0       | 100000   | 100000   |
|                 | BuFAU     | 178.79 | 21.72   | 207.78   | 256.75  | 100000  | 0       | 100000   | 100000   |
|                 | greedy    | 178.79 | 21.72   | 207.78   | 256.75  | 100000  | 0       | 100000   | 100000   |
|                 | CBM       | 178.79 | 21.72   | 207.78   | 256.75  | 100000  | 0       | 100000   | 100000   |

Table 3: Additional statistics of the empirical evaluation in Figure 4. All statistics are measured after $T = 100,000$ time steps over 1,000 different seeds. All experiments use $\epsilon(t) = \frac{1}{\ln t}$. std is the standard deviation, 90% represents the 90th percentile and max is the maximum.

| Bandit Instance | Algorithm | Regret | Queries |
|-----------------|-----------|--------|---------|
|                 | mean      | std    | 90%     | max      | mean    | std    | 90%     | max      |
| 5-arms, unique optimal, values: $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$ | BuFALU    | 551.43 | 73.82   | 644.08   | 780.25  | 2881.58 | 373.37  | 3336.5   | 4072     |
|                 | BuFAU     | 200.47 | 21.92   | 228.75   | 276.25  | 9958.88 | 87.68   | 10072.2  | 10262    |
|                 | greedy    | 236.37 | 24.63   | 269.02   | 311.5   | 45785  | 0       | 45785    | 45785    |
|                 | CBM       | 200.52 | 21.94   | 228.82   | 276.25  | 9959.06 | 87.76   | 10072.3  | 10262    |
| 5-arms, multiple optimal, values: $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$ | BuFALU    | 150.97 | 17.9    | 173.78   | 210     | 16349.9| 1762.22 | 18277.3  | 18954    |
|                 | BuFAU     | 147.59 | 18.04   | 170      | 235.5   | 16811.8| 1387.87 | 18504.5  | 19022    |
|                 | greedy    | 167.42 | 20.59   | 194.02   | 238.25  | 45785  | 0       | 45785    | 45785    |
|                 | CBM       | 147.57 | 17.99   | 170.5    | 235.5   | 16827.3 | 1390.31 | 18523.1  | 19022    |
C.2.2 Testing the Effect of the Total Budget

We now test the effect of the total budget on the regret and number of feedback queries. We do so problems with Bernoulli arms, optimal arm $\mu^* = 0.5$, a single suboptimal arm $\mu_a = 0.25$ and either a unique or two optimal arms (namely, two and three-armed problems, respectively). The time horizon in the simulations is $T = 1,000$ and results were averaged over 1,000 seeds. To simulate the effect of the total budget, we allocated a fixed budget $B$ throughout the interactions (i.e., set $\epsilon(t) = \sqrt{\frac{6K \ln t}{B}}$) and measured both the regret and number of queries at the last time step, as a function of the (fractional) allocated budget $B/T$. The results are depicted in Figure 5 and error bars represent a variation of one standard deviation.

Interestingly, the regret first decreases, achieves a minimum when $B/T \approx 0.08$, and then starts increasing. We believe that this is since the standard UCB bounds are a little loose and can be replaced by [Garivier and Cappé, 2011]:

$$UCB_t(a) = \hat{\mu}_t(a) + \sqrt{\frac{\ln t + 3 \ln \ln t}{2n_{t-1}^a(a)}}.$$ 

Therefore, the algorithm suffers from some over-exploration. Then, the local minimum of the regret allocates enough budget to adequately explore and then forces the algorithm the exploit, which leads to a better exploration-exploitation tradeoff.

Asides from this phenomenon, the graph behaves as could be expected. As we previously saw, all algorithms achieve roughly the same performance in the presence of multiple optimal arms. When the optimal arm is unique, BuFALU is substantially fewer budget queries but suffers from a small regret degradation (up to a factor of 4). Notably, in the small-budget regime, BuFALU achieves similar regret to all other baselines.

![Figure 5: Regret and Queries as a function of the allocated budget in two and three armed bandit problems.](image-url)