Understanding the Gaps in Satisficing Bandits

Chloé Rouyer
Montanuniversität Leoben, Austria
chloerouyer.ml@gmail.com

Ronald Ortner
Montanuniversität Leoben, Austria
rortner@unileoben.ac.at

Peter Auer
Montanuniversität Leoben, Austria
auer@unileoben.ac.at

Abstract

In this work, we consider a variation of the stochastic multi-armed bandit problem in which the learner is not necessarily trying to compete with the best arm, whose performance is not known ahead of time, but is satisfied with playing any arm that performs above a known satisficing threshold S. Michel et al. (2023) considered as respective performance measure the *satisficing regret*, that scales in terms of the gaps between the expected performance of an insufficient arm and the threshold S, rather than in terms of its gap with the best arm. While Michel et al. propose an algorithm that achieves time-independent satisficing regret, their results suffer when arms are too close to the threshold. Is this dependency unavoidable? The first contribution of our work is to provide an alternative and more general lower bound for the K-armed satisficing bandit problem, which highlights how the position of the threshold compared to the arms affects the bound. Then, we introduce an algorithm robust against unbalanced gaps, which enjoys a nearly matching time-independent upper bound. We also propose an alternative definition of the satisficing regret, which might be better tailored to measure algorithm performance in these difficult instances and derive a lower bound for this regret. Finally, we include experiments to compare these different regret measures and our proposed algorithms empirically.

1 Introduction

The stochastic multi-armed bandit problem is one of the most fundamental frameworks when it comes to studying the trade-off between exploration and exploitation in sequential decision-making (Thompson, 1933; Auer, 2002; Bubeck & Cesa-Bianchi, 2012; Slivkins, 2019; Lattimore & Svepesvári, 2020). In the standard version of the problem, the learner's goal is to achieve performance as close to the performance of the best arm in expectation as possible. However in practice when faced with repetitive tasks, it is very common to focus on achieving a sufficient performance rather than an optimal one. One way to formulate this problem is by considering a satisficing threshold and modifying the objective of the learner to be satisfied with playing any arm that has an expected performance above the threshold. Many problems can fit this framework such as problems where a learner has to stay under a given budget when commissioning work from service providers, dynamic resource consumption where electric vehicles have to charge at night when electricity is cheap while ensuring a certain battery charge rate in the morning, or the average student aiming to get a passing grade rather than the best grade.

This notion of a satisficing threshold has been studied in the literature, mainly in the bandit framework, but a recent work generalized these results towards reinforcement learning. Early evaluations of this notion of satisfiability in multi-armed bandits include the works of Kohno & Takahashi (2017);

17th European Workshop on Reinforcement Learning (EWRL 2024).

Tamatsukuri & Takahashi (2019), which show that if the satisficing threshold, called aspiration level, is placed between the best arm and the second best arm, then it is possible to achieve a time-independent regret as the notions of satisficing and optimizing converge.

Recently, Michel et al. (2023) generalized this result and considered as performance measure the *satisficing pseudo-regret*. This measure evaluates the satisficing performance by taking into account only the gaps between 'insufficient arms' and the satisficing threshold S rather than the gaps between these insufficient arms and the best arm. The main downside of this approach is that this measure can only be optimized in the realizable setting, meaning that we have to know ahead of time that there exists a satisficing arm. Otherwise, if all the arms are below the threshold then the satisficing pseudo-regret would necessarily scale linearly with the time horizon T. To counter this issue, Michel et al. (2023) propose an algorithm that enjoys time-independent satisficing regret in the realizable case while also ensuring a rate-optimal logarithmic upper bound on the pseudo-regret in the non-realizable case. In the realizable case, the algorithm of Michel et al. achieves an $O\left(\sum_{i:\mu_i<S} \widetilde{\Delta}_i + \frac{1}{\widetilde{\Delta}_i} + \frac{\widetilde{\Delta}_i}{\Delta_*^2}\right)$ satisficing regret guarantee, where $\widetilde{\Delta}_i = S - \mu_i$ is the insufficiency gap and $\Delta_* = \max_j \mu_j - S$ is the gap between the best arm and the threshold. This bound is time-independent, but it also scales inversely proportionally to both $\widetilde{\Delta}_i$ and Δ_*. This means that if either of these gaps is small, for example of order $\frac{1}{T}$, then this satisficing regret bound would be vacuous whereas the standard pseudo-regret bound is still of order $O\left(\sum_{i:\Delta_i>0} \frac{\log T}{\Delta_i}\right)$, and the standard suboptimality gaps Δ_i can be large even when Δ_* or $\widetilde{\Delta}_i$ is not. Michel et al. (2023) also propose a lower bound that scales as $\Omega\left(\sum_{i:\widetilde{\Delta}_i>0} \frac{1}{\widetilde{\Delta}_i}\right)$. In its construction, this lower bound assumes that the gaps $\widetilde{\Delta}_i = \Delta_*$ are the same for all arms i below the threshold.

The first contribution of our work is a novel derivation of this lower bound that specifically focuses on cases where the gaps can be unbalanced. We consider a satisficing bandit problem in the realizable case and show a lower bound on the regret of $\Omega\left(\sum_{i:\widetilde{\Delta}_i>0} \frac{\widetilde{\Delta}_i}{\Delta_i^2}\right)$.

Then, we propose an alternative definition of the satisficing regret which offers a smoother transition between the non-realizable case, the realizable case with a unique sufficient arm, and the realizable case with multiple sufficient arms. This regret definition measures the insufficiency gap in terms of the gap between a bad arm and the worst of the sufficient arms rather than the gap to the threshold S. Focusing on the two-armed case, we derive a lower bound on this regret measure of order $\Omega\left(\frac{\log(\frac{\Delta}{\Delta_*})}{\Delta}\right)$ which better captures the challenge of unbalanced gaps. We propose a novel algorithm, `uncertain-UCB`, and show that it enjoys near-optimal time-independent guarantees with respect to this novel satisficing regret measure. The proof of this bound is shown confined to the two-armed setting. However, we present experiments that indicate that the analysis should generalize. These experiments compare the performance of `uncertain-UCB` and `SAT-UCB` (Michel et al., 2023) against the baseline `UCB1` (Auer, 2002) in terms of the standard pseudo-regret, the satisficing regret of Michel et al. (2023), and our novel satisficing regret.

1.1 Related Literature

Reverdy et al. (2017) introduced the notion of satisficing bandits in the Bayesian setting. Following the same approach in a frequentist setting, Hüyük & Tekin (2021) propose a reduction of the satisficing bandit setting to bandits with lexicographically ordered objectives. In this setting, they show that their algorithm achieves a satisficing pseudo-regret bound of order $O\left(\sum_{i:\widetilde{\Delta}_i>0} \frac{\log \frac{1}{\Delta_i}}{\widetilde{\Delta}_i}\right)$. We also note that similar results can be obtained when running the algorithm of Garivier et al. (2019) using the satisficing threshold instead of the optimal mean reward to tune the algorithm.

Recently, the first generalization of satisficing bandits to reinforcement learning has been made by Hajiabolhassan & Ortner (2023). They consider the reinforcement learning problem in a communicating MDP with finite state and action spaces and propose the `SAT-RL` algorithm, which achieves time-independent satisficing regret when a satisficing policy exists.

For the lower bounds, Bubeck et al. (2013, Theorem 5) and Garivier et al. (2019, Theorem 7) consider a two-player bandit game where the mean of the best arm μ^* is known and show that an asymptotic

dependency on $\frac{1}{\Delta}$ is unavoidable, even when the gaps are unknown. We generalize this result to K-armed bandits problems where the learner knows the satisficing threshold S, which includes knowledge of μ^* in a regular bandit game as a special case, where $S = \mu^*$.

2 Problem Setting and Notation

We consider a bandit problem with K arms played over the course of T rounds. Following the setting proposed by Michel et al. (2023), we assume that the environment generates rewards by sampling them from the class of sub-Gaussian distributions. We recall that for a sub-Gaussian distribution ν with mean μ and empirical estimate $\hat{\mu}_n$ computed from n samples it holds for any $\varepsilon > 0$:

$$\begin{aligned} \mathbb{P}\left[\hat{\mu}_n > \mu + \varepsilon\right] &\leq e^{-n\varepsilon^2/2}, \\ \mathbb{P}\left[\hat{\mu}_n < \mu - \varepsilon\right] &\leq e^{-n\varepsilon^2/2}. \end{aligned} \tag{1}$$

We note that any bounded distribution is sub-Gaussian and refer to Vershynin (2018) for further properties of sub-Gaussian distributions.

In the classic multi-armed bandit problem, the performance of the learner after T rounds is evaluated in terms of the pseudo-regret, defined as:

$$R_T := \sum_{i:\Delta_i > 0} \Delta_i \, \mathbb{E}\left[N_i(T)\right], \tag{2}$$

where $\Delta_i := \mu^* - \mu_i$ is the gap between the optimal mean reward and the mean reward of arm i, and $N_i(T)$ denotes the number of times arm i is played in T rounds. In this work, we consider variations of the bandit problem where the focus is placed on achieving a sufficient performance rather than an optimal one. In Section 3, we focus on the setting as defined by Michel et al. (2023), where the learner knows the satisficing threshold S ahead of time, and performance is evaluated in terms of the satisficing pseudo-regret (short S-pseudo-regret), defined as:

$$\mathcal{R}_T^S := \sum_{i:\widetilde{\Delta}_i > 0} \widetilde{\Delta}_i \, \mathbb{E}\left[N_i(T)\right], \tag{3}$$

where $\widetilde{\Delta}_i := \max\left\{S - \mu_i, 0\right\}$ measures the gap between an insufficient arm i and the threshold S. Furthermore, we also define $\Delta_* := \mu^* - S$ as a shorthand for the gap between the mean reward of the best arm and the threshold S.

Remark 2.1. As above, in the following we use the terms *sufficient* (resp. *insufficient*) to denote whether the expected reward of an arm is above (resp. below) the threshold S. We also use the term *unbalanced gaps* to refer to a problem setting where the order of magnitude of Δ_* or $\widetilde{\Delta}_i$ is very small compared to Δ_i. We also use the following notations: $\mathbb{P}_\nu$ (resp. $\mathbb{E}_\nu$) defines the probability (respectively the expectation) of an event happening in the bandit problem ν.

Remark 2.2. It is easy to see that when all the arms are below the threshold S and $\min_i\left(S - \mu_i\right) = \Delta > 0$, then

$$\mathcal{R}_T^S \geq \Delta T.$$

This means that studying the S-pseudo-regret only makes sense when there exists at least one sufficient arm.

Definition 2.3. A bandit problem is *S-realizable* if there exists an arm $i \in [K]$ such that $\mu_i > S$.

3 A Lower Bound for the Satisficing Pseudo-Regret

In this section, we focus on the setting proposed by Michel et al. (2023) and derive a generalized lower bound that holds for K arms and does not assume that the gaps Δ_* and $\widetilde{\Delta}_i$ are equal. This result highlights that the algorithms of Michel et al. enjoy tighter guarantees than initially anticipated, even though there is still a gap when the rewards are unbalanced.

We introduce the following definitions to formalize this result.

Definition 3.1. An algorithm is considered *stable* if for any bandit problem ν and arms $i, j \in [K]$ with $\mu_i = \mu_j$ it holds for all t:

$$\mathbb{E}\left[N_i(t)\right] = \mathbb{E}\left[N_j(t)\right].$$

This constraint is quite mild as most algorithms are stable by permutation over the arms. We also use another definition, which ensures that insufficient arms are only played a constant number of times.

Definition 3.2. An algorithm is *S-satisficing* if for any sub-Gaussian bandit problem ν there exists a constant C such that:

$$\lim_{T\to\infty} R_T^S \leq C.$$

This constant C can be problem-dependent and in particular may scale with Δ_i or $\widetilde{\Delta}_i$.

With these definitions, we can move on to state the following result, whose proof follows Garivier et al. (2019, Theorem 7).

Theorem 3.3. *For any stable S-satisficing algorithm, there exists an S-realizable bandit problem with sub-Gaussian distributions for which this algorithm admits the following lower bound:*

$$\lim_{T\to\infty} \mathcal{R}_T^S \geq \sum_{i:\widetilde{\Delta}_i>0} \frac{\widetilde{\Delta}_i}{\Delta_i^2}.$$

Proof of Theorem 3.3. We consider a bandit problem ν with a unique sufficient and optimal arm i^*, where all the arms follow Gaussian distributions: for all $i \in [K]$, $\nu_i = \mathcal{N}(\mu_i, 1)$. In this proof, we use $\mathrm{KL}(P, Q)$ to denote the standard Kullback–Leibler divergence between distributions P and Q and $\mathrm{kl}(p, q)$ to denote the Kullback–Leibler divergence between two Bernoulli distributions with respective parameters p and q.

Following the definition of the S-pseudo-regret in Equation (3), we want to lower bound the number of times each suboptimal i arm is played in ν. To do so, for each such arm i we consider another bandit problem ν' defined such that $\forall j \neq i$, $\nu'_j = \nu_j$ and $\nu'_i = \nu_{i^*}$. This choice of ν' fulfills two important properties. First, the two problems differ only on arm i, meaning that:

$$\sum_{j=1}^{K} \mathbb{E}_\nu[N_j(T)]\,\mathrm{KL}(\nu_j, \nu'_j) = \mathbb{E}_\nu[N_i(T)]\,\mathrm{KL}(\nu_i, \nu'_i) \geq \mathrm{kl}(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z]), \tag{4}$$

where the last step follows from Garivier et al. (2019, Equation 6) and holds for any bandit problems ν, ν' and $\sigma(I_T)$-measurable random variables Z with values in $[0, 1]$, where I_T contains all the past information obtained by the algorithm up to round T. This holds in particular when picking $Z = N_i(T)/T$, as we will do in the following.

As $\nu_i = \mathcal{N}(\mu_i, 1)$ and $\nu'_i = \mathcal{N}(\mu_{i^*}, 1)$, we can ensure that $\mathrm{KL}(\nu_i, \nu'_i) = \frac{\Delta_i^2}{2}$. Combining these two statements with Equation (4), we obtain:

$$\frac{\Delta_i^2}{2}\mathbb{E}_\nu[N_i(T)] = \mathbb{E}_\nu[N_i(T)]\,\mathrm{KL}(\nu_i, \nu'_i) \geq \mathrm{kl}\left(\frac{\mathbb{E}_\nu[N_i(T)]}{T}, \frac{\mathbb{E}_{\nu'}[N_i(T)]}{T}\right). \tag{5}$$

Then, in setting ν', we know that both arm i and i^* have the same distribution, meaning that Definition 3.1 ensures that $\mathbb{E}_{\nu'}[N_i(T)] = \mathbb{E}_{\nu'}[N_{i^*}(T)]$. As these arms are the two only sufficient arms and our algorithm is S-satisficing, all the other arms are played at most a constant number of times and we have:

$$\lim_{T\to\infty} \frac{\mathbb{E}_{\nu'}[N_i(T)]}{T} = \frac{1}{2}. \tag{6}$$

Rearranging Equation (5) and taking the limit on T gives:

$$\begin{aligned}
\lim_{T\to\infty} \mathbb{E}_\nu[N_i(T)] &\geq \frac{2}{\Delta_i^2} \lim_{T\to\infty} \mathrm{kl}\left(\frac{\mathbb{E}_\nu[N_i(T)]}{T}, \frac{\mathbb{E}_{\nu'}[N_i(T)]}{T}\right) \\
&\geq \frac{4}{\Delta_i^2} \lim_{T\to\infty} \left(\frac{\mathbb{E}_\nu[N_i(T)]}{T} - \frac{\mathbb{E}_{\nu'}[N_i(T)]}{T}\right)^2 \\
&\geq \frac{4}{\Delta_i^2} \left(\lim_{T\to\infty} \frac{\mathbb{E}_\nu[N_i(T)]}{T} - \lim_{T\to\infty} \frac{\mathbb{E}_{\nu'}[N_i(T)]}{T}\right)^2 \\
&\geq \frac{4}{\Delta_i^2} \left(0 - \frac{1}{2}\right)^2 \\
&= \frac{1}{\Delta_i^2},
\end{aligned}$$

where the first step reorganizes Equation (5), the second step applies Pinsker's inequality, which ensures that for any $p, q \in [0,1]$, $\mathrm{kl}(p,q) \geq 2(p-q)^2$. Then, the third step uses the properties of the limit: as both $\lim_{T\to\infty} \frac{\mathbb{E}_\nu[N_i(T)]}{T} = 0$ and $\lim_{T\to\infty} \frac{\mathbb{E}_{\nu'}[N_i(T)]}{T} = \frac{1}{2}$ are finite, and the square function is continuous, we can move the limits inside the square function.

Repeating this bound for all suboptimal arms finishes the proof. □

4 A Novel Definition of the Satisficing Pseudo-Regret

In this section, we consider a novel variation of the satisficing pseudo-regret, which measures the gaps between the smallest sufficient arm $\bar{\mu} = \min_{i:\mu_i > S} \mu_i$ and the insufficient arms:

$$\bar{\mathcal{R}}_T^S = \sum_{i:\tilde{\Delta}_i > 0} (\bar{\mu} - \mu_i)\, \mathbb{E}\left[N_i(T)\right]. \tag{7}$$

We note that by construction, this novel satisficing regret is always larger than the S-pseudo-regret of Michel et al. (2023), and no larger than the standard pseudo-regret. It offers a smooth transition between the realizable case and the non-realizable case. This definition is particularly interesting in cases where there exists a single sufficient arm, as then $\bar{\mu} = \mu^*$, and for all insufficient arms j we have $\bar{\mu} - \mu_j = \Delta_j$ and $\bar{\mathcal{R}}_T^S = \mathcal{R}_T$, meaning that the novel satisficing regret matches the standard pseudo-regret. Furthermore, when the gaps are balanced, both definitions of satisficing regret are almost equivalent and only differ by a small constant factor.

4.1 Lower bound

In this section, we present a lower bound for this novel satisficing pseudo-regret.

Theorem 4.1 (Lower Bound for Two Arms)**.** *For any S-satisficing algorithm, there exist two-armed bandit problems ν and ν' and a time horizon T such that for all $t \geq T$:*

$$\max\left(\bar{\mathcal{R}}_t^S(\nu), \bar{\mathcal{R}}_t^S(\nu')\right) \geq \frac{\max\left\{1, \log \frac{\Delta}{\Delta_*}\right\}}{8\Delta},$$

where Δ_* is the gap between the best arm and the threshold and Δ is the sub-optimality gap between the two arms.

Proof. Consider a two-armed bandit problem ν and a threshold S where the arms follow Gaussian distributions such that $\nu_i = \mathcal{N}(\mu_i, 1)$ for $i = 1, 2$, and the means fulfill $\mu_1 = S + \Delta_*$ and $\mu_2 = \mu_1 - \Delta < S$.

We also consider a modified bandit problem ν' such that $\mu_1' = S - \Delta_*$ and $\mu_2' = \mu_1' + \Delta$.

We want to show that for T sufficiently large, no algorithm can achieve low regret on both instances simultaneously. Let T and $\alpha > 0$ be constants that we will set later.

For any instance of a two-player bandit game and any learning algorithm, after $\frac{\log\alpha}{4\Delta^2} + \frac{1}{4\Delta_*^2} + 1$ rounds, exactly one of these conditions hold:

$$\underbrace{N_1\left(\frac{\log\alpha}{4\Delta^2} + \frac{1}{4\Delta_*^2}\right) > \frac{1}{4\Delta_*^2}}_{\text{Condition 1}} \quad \text{or} \quad \underbrace{N_2\left(\frac{\log\alpha}{4\Delta^2} + \frac{1}{4\Delta_*^2}\right) > \frac{\log\alpha}{4\Delta^2}}_{\text{Condition 2}}.$$

If the second condition holds for the bandit game ν, then $\bar{\mathcal{R}}_T^S(\nu) > \frac{\log\alpha}{4\Delta}$ follows and our result trivially holds. Thus we want to show that when the first condition holds we cannot obtain low regret for the bandit game ν'.

Consider a time horizon T such that $N_1(T) = \frac{1}{4\Delta_*^2}$ and $N_2(T) < \frac{\log\alpha}{4\Delta^2}$. Under Condition 1, we know that such T exists.

We recall that $\forall\gamma > 0$, $\forall\,\bar{\nu} \in \{\nu_1, \nu_2\}$ and $\forall i \in \{1, 2\}$,

$$\mathbb{E}_{\bar{\nu}}[N_i(T)] \geq \gamma\,\mathbb{P}_{\bar{\nu}}[N_i(T) > \gamma]. \tag{8}$$

Following Theorem 14.2 of Lattimore & Svepesvári (2020), we have:

$$\begin{aligned}
\mathbb{P}_\nu[N_1(T) < \gamma] + \mathbb{P}_{\nu'}[N_1(T) \geq \gamma] &\geq \frac{1}{2}\exp\left(-\mathrm{KL}(\nu, \nu')\right) \\
&= \frac{1}{2}\exp\left(-\left(\mathbb{E}_\nu[N_1(T)]\frac{4\Delta_*^2}{2} + \mathbb{E}_{\nu'}[N_2(T)]\frac{4\Delta^2}{2}\right)\right) \\
&\geq \frac{1}{2}\exp\left(-\left(\frac{1}{2} + \frac{1}{2}\log\alpha\right)\right) \\
&\geq \frac{1}{2}\exp\left(-\log\alpha\right) \\
&= \frac{1}{2\alpha},
\end{aligned}$$

where the second step follows from Lemma 15.1 of Lattimore & Svepesvári (2020) and Proposition A.2 in the appendix, where we note that the gaps between the two distributions are $2\Delta_*$ and $2\Delta - 2\Delta_* \geq 2\Delta$, respectively. Then, the third step uses the value of $N_1(T)$ we picked, and as $\exp(-x)$ is a decreasing function, we can lower bound the second sum by picking an upper bound for $\mathbb{E}_{\nu'}[N_2(T)] \geq \frac{\log\alpha}{4\Delta^2}$. The fourth step uses that $\exp(-x)$ is a decreasing function and upper bounds $\frac{1}{2} + \frac{1}{2}\log\alpha$ by $\log\alpha$, which holds for any α such that $\log\alpha \geq 1$.

We pick $\gamma = \frac{1}{4\Delta_*^2}$ so that $\mathbb{P}_\nu[N_1(T) < \frac{1}{4\Delta_*^2}] = 0$ by definition of T, and we deduce that $\mathbb{P}_{\nu'}[N_1(T) \geq \frac{1}{4\Delta_*^2}] \geq \frac{1}{2\alpha}$. Combining this result with Equation (8), we have:

$$\bar{\mathcal{R}}_T^S(\nu') = \Delta\mathbb{E}_{\nu'}[N_1(T)] \geq \frac{1}{2\alpha}\frac{\Delta}{4\Delta_*^2}.$$

This result, combined with our initial assumption, ensures that for all α such that $\log\alpha \geq 1$:

$$\max\left(\bar{\mathcal{R}}_T^S(\nu), \bar{\mathcal{R}}_T^S(\nu')\right) \geq \frac{1}{2}\left(\bar{\mathcal{R}}_T^S(\nu) + \bar{\mathcal{R}}_T^S(\nu')\right) \geq \frac{1}{8}\left(\frac{\log\alpha}{\Delta} + \frac{1}{2\alpha}\frac{\Delta}{\Delta_*^2}\right). \tag{9}$$

Then, we note that choosing α such that $2\alpha\log\alpha = \frac{\Delta^2}{\Delta_*^2}$ would maximize the right hand side of Equation (9). Using our condition $\log\alpha \geq 1$, we deduce that $\frac{\Delta^2}{\Delta_*^2} = \alpha\log\alpha \leq \alpha^2$. This means that choosing α such that $\log\alpha = \max\left\{1, \log\frac{\Delta}{\Delta_*}\right\}$ both fulfills the condition $\log\alpha \geq 1$ and is close to the optimal choice of α without the constraint $\log\alpha \geq 1$. We deduce that:

$$\max\left(\bar{\mathcal{R}}_T^S(\nu), \bar{\mathcal{R}}_T^S(\nu')\right) \geq \frac{\max\left\{1, \log\frac{\Delta}{\Delta_*}\right\}}{8\Delta},$$

which finishes the proof. □

4.2 Upper Bound

We propose the novel `uncertain-UCB` algorithm shown as Algorithm 1. Following the structure of the `UCB1` algorithm, the key difference between both methods lies in the construction of the upper confidence bounds: instead of using the current step count t in the confidence intervals (cf. Equation 12 below), we use a proxy $n_0(t)$, called the number of uncertain rounds. This value, defined in Equation (10), presents the interesting characteristic of only increasing when the learner plays an arm whose empirical estimate is not sufficiently far above the threshold S. This means that as long as the learner cannot be reasonably certain that the arms that are played are satisficing, $n_0(t)$ keeps on increasing and remains close to t. During this phase, the algorithm behaves like the classic `UCB1`. Once the algorithm plays arms that are sufficiently far above the threshold, the number of uncertain rounds $n_0(t)$ stops increasing and the upper confidence bounds of the arms stop growing. This approach slows the exploration of all arms, and in particular of insufficient arms, which allows the algorithm to achieve a time-independent satisficing regret.

Algorithm 1: `uncertain-UCB`

Input: number of arms K, satisficing threshold S.
Play each arm once, i.e., for time steps $t = 1, \dots, K$ choose $i(t) = t$.
for $t = K + 1, \dots$ **do**
Compute

$$n_0(t) = \sum_{\tau<t} \mathbb{I}\left[\hat{\mu}_{i(\tau)}(\tau) < S + \sqrt{C_2 \frac{\log n_{i(\tau)}(\tau)}{n_{i(\tau)}(\tau)}}\right], \tag{10}$$

$$\forall i \in [K]: \quad u_i(t) = \hat{\mu}_i(t) + \sqrt{C_1 \frac{\log n_i(t) + \log n_0(t)}{n_i(t)}}, \tag{11}$$

where $\hat{\mu}_i(t)$ is the empirical estimate and $n_i(t)$ the count of arm i at time t.
Play $i(t) = \operatorname{argmax}_{i\in[K]} u_i(t)$.
Observe $r_{i(t)}(t)$ and update the empirical estimate $\hat{\mu}_{i(t)}$.
end for

Theorem 4.2. *For any S-realizable bandit problem with means $\mu_i \in [0, 1]$, `uncertain-UCB` with $C_1 \geq 4$ and $C_2 \geq 6$ satisfies for all $T > 1$,*

$$\bar{\mathcal{R}}_T^S \leq O\left(\frac{1}{\Delta_2}\left(\log\frac{1}{\Delta_*} + \log\frac{1}{\Delta_2}\right) + \sum_{\kappa\geq 4} \kappa\Delta_*^{2\kappa-3} \log\frac{1}{\Delta_*}\right).$$

The sum in the bound converges for all $\Delta_* < 1$, and empirical experiments presented in Section 5 show that small values of Δ_* don't seem to affect the behavior of the algorithm noticeably, which suggests that the algorithm obtains a smaller dependency on $\frac{1}{\Delta_*}$. The proof of Theorem 4.2 is given in Appendix B.

5 Experiments

In this section, we provide some experiments to compare the `uncertain-UCB` algorithm against the `SAT-UCB` algorithm. Here `uncertain-UCB` is run using $C_1 = 4$ and $C_2 = 6$. We also use the `UCB1` algorithm as defined by Auer (2002) using index values

$$u_i(t) = \hat{\mu}_i(t) + \sqrt{2\frac{\log t}{n_i(t)}} \tag{12}$$

in these experiments to serve as a baseline.

Our experiments are run with eight arms and are repeated 10 times. The rewards are sampled from Bernoulli distributions with respective means $[1, 0.875, 0.75, 0.625, 0.5, 0.375, 0.25, 0.125]$.

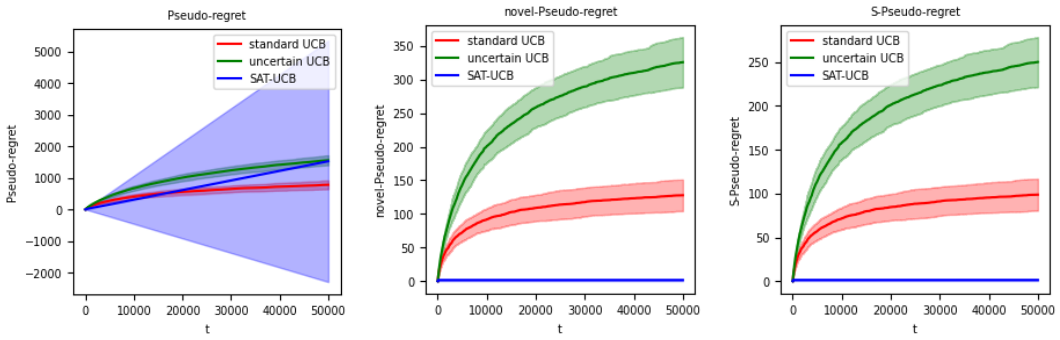

Figure 1: First experiment, with four sufficient arms and four insufficient arms and evenly balanced gaps. The plots are sorted from the largest regret measure to the smallest.

For each of the algorithms, we plot the empirical regret, the empirical satisficing pseudo-regret, and the empirical novel pseudo-regret. The plots show the average across the 10 runs, and the shaded areas represent the standard deviation across these runs.

In the first experiment presented in Figure 1, the threshold is the mean between the fourth and fifth best arms: we have four sufficient arms and four insufficient arms, and the gaps are evenly balanced around the threshold. As expected, both notions of satisficing regret behave similarly. We observe that `SAT-UCB` has an interesting behavior: when evaluating the pseudo-regret, this algorithm suffers linear regret, whereas it enjoys near-constant satisficing regret. This suggests that the algorithm identifies and plays satisficing arms fast, but does not perform further exploration to find a better, and possibly optimal arm. As expected, `UCB1` behaves similarly across all regret measures and the same holds for and `uncertain-UCB`. It is worth noting that `UCB1` outperforms `uncertain-UCB`, which we conjecture is due to the term $4(\log n_i(t) + \log n_0(t))$ being large compared to $2 \log t$ in the initial rounds, which forces more exploration.

We then performed a second experiment highlighting the limits of the SAT-UCB algorithm. We consider a problem with the same arms as before, but instead of having S be centered in between four sufficient and four insufficient arms, we consider two new cases: one where S is just below the best arm, and one where the threshold is just below the third-best arm. In both cases, 'just below' defines a gap of 0.001 between the threshold and the corresponding arm, which is orders of magnitude smaller than the gaps between the arms. As the three performance measures provided similar results, we only display the results for the novel satisficing regret in Figure 2. These results highlight that `SAT-UCB` scales with Δ_*^{-1}, and thus its performance appears linear as the time horizon needed to learn Δ_*^{-1} is large compared to the time horizon to distinguish between the arms without use of the threshold. This experiment highlights that arms being close to the threshold, and in particular of the best arm are especially challenging for the SAT-UCB algorithm and show use-cases where the robustness of algorithms such as `uncertain-UCB` is valuable.

6 Discussion

We characterized the complexity of the satisficing bandit problem by deriving new lower bounds for the problem, highlighting that satisficing can be a difficult problem to optimize for when the gaps are unbalanced. We proposed a novel definition of satisficing regret and showed both upper and lower bounds in the two-armed case. Experiments highlight that our new `uncertain-UCB` algorithm should perform well in the standard K-arms bandits problem.

The main direction for future work is to generalize the analysis of this algorithm to more than two arms.

Acknowledgements

This work was supported by the Austrian Science Fund (FWF): TAI 590-N.

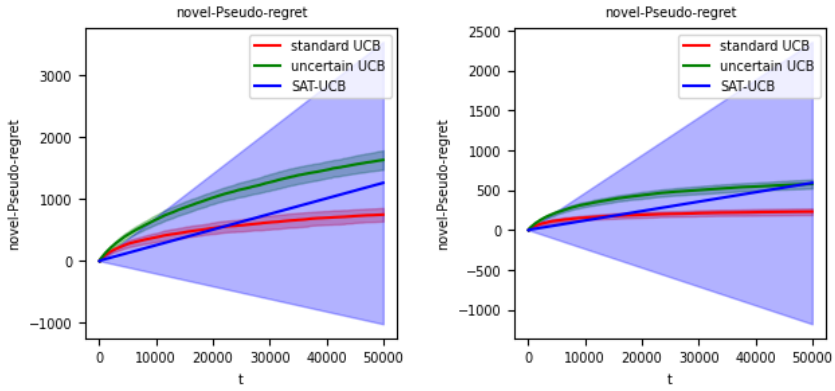

Figure 2: For the third experiment we considered two settings. The first (shown on the left) has a single sufficient arm and seven insufficient arms, where the threshold is just below the best arm (with a gap of 0.001 with that arm). The second setting has (shown on the right) has three sufficient arms and five insufficient arms, where the threshold is very close to the third best arm (with a gap of 0.001 between the third-best arm and the threshold, and $\Delta_* = 0.25 + 0.001$). In both examples, the gaps are unbalanced around S. Both results are plotted in terms of the novel satisficing regret.

References

Auer, P. Using confidence bounds for exploration-exploitation trade-offs. *Journal of Machine Learning Research*, 3, 2002.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5, 2012.

Bubeck, S., Perchet, V., and Rigollet, P. Bounded regret in stochastic multi-armed bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2013.

Chatzigeorgiou, I. Bounds on the lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 2013.

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 2019.

Hajiabolhassan, H. and Ortner, R. Online regret bounds for satisficing in MDPs. In *Sixteenth European Workshop on Reinforcement Learning*, 2023.

Hüyük, A. and Tekin, C. Multi-objective multi-armed bandit with lexicographically ordered and satisficing objectives. In *Machine Learning, Springer*, 2021. doi: https://doi.org/10.1007/s10994-021-05956-1.

Kohno, Y. and Takahashi, T. A cognitive satisficing strategy for bandit problems. *International Journal of Parallel, Emergent and Distributed Systems*, 2017. doi: 10.1080/17445760.2015.1075531.

Lattimore, T. and Svepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

Michel, T., Hajiabolhassan, H., and Ortner, R. Regret bounds for satisficing in multi-armed bandit problems. *Transactions on Machine Learning Research*, 2023.

Reverdy, P., Srivastava, V., and Leonard, N. E. Satisficing in multi-armed bandit problems. *IEEE Transactions on Automatic Control*, 2017.

Rouyer, C., van der Hoeven, D., Cesa-Bianchi, N., and Seldin, Y. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Slivkins, A. Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 12, 2019.

Tamatsukuri, A. and Takahashi, T. Guaranteed satisficing and finite regret: Analysis of a cognitive satisficing value function. *Biosystems*, 2019.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25, 1933.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. ISBN 978-1-108-41519-4.

A Useful Results

We recall some straightforward results useful in the remainder of the paper.

Proposition A.1. *Let a_n and b_n be two functions of n such that for all $n > 0$, we have $a_n, b_n \geq 0$. Then we have:*

1. $\sum_{n>0} e^{a_n - b_n} \leq \sum_{n>0} e^{a_n}$,
2. $\sum_{n_1,n_2>0} e^{a_{n_1}+b_{n_2}} \leq \sum_{n_1>0} e^{a_{n_1}} \sum_{n_2>0} e^{b_{n_2}}$

This result is standard and stated here for completeness.

Proposition A.2. *Consider $\nu_1 = \mathcal{N}(\mu_1, 1)$ and $\nu_2 = \mathcal{N}(\mu_2, 1)$ two uni-variate Gaussian distributions with unit variance, such that $|\mu_1 - \mu_2| = \Delta$. Then,*

$$\mathrm{KL}(\nu_1, \nu_2) = \frac{\Delta^2}{2}.$$

B Analysis of the `uncertain-UCB` Algorithm

While the full proof only holds in the two-armed case due to limitations in the analysis of Lemma B.5 below, we highlight that some parts of the analysis are given for the general K-armed setting.

We consider an S-realizable bandit problem and assume without loss of generality that $\mu_1 > S > \mu_2$ in the two-armed case. We also decompose the number of uncertain rounds and set for any arm j:

$$n_{0,j}(t) := \sum_{\tau<t} \mathbb{I}\left[I_\tau = j, \hat{\mu}_{i(\tau)}(\tau) < S + \sqrt{C_2 \frac{\log n_{i(\tau)}(\tau)}{n_{i(\tau)}(\tau)}}\right]. \tag{13}$$

Remark B.1. When needed, we use the notation $\hat{\mu}_{i,n}$ to denote the empirical mean of any arm i built from n samples rather than the $\hat{\mu}_i(t)$ notation, which represents the empirical mean of arm i at round t.

The key difference between the standard analysis of `UCB1` and the present analysis of `uncertain-UCB` comes from handling the uncertain rounds. To contribute to the count of uncertain rounds, an arm has to be played and remain uncertain, and these two cases cannot happen simultaneously too often. For the arms closest to the best arm, we show that with high probability, the number of times that arm is uncertain is bounded by a constant.

Lemma B.2. *When Algorithm 1 is run with $C_2 \geq 6$, the probability that sufficient arm j contributes to more than n rounds of the uncertain rounds is bounded as:*

$$\mathbb{P}\left[n_{0,j}(T+1) > n\right] \leq \frac{2\exp\left(-(n-1)\bar{\Delta}_j^2/2\right)}{\bar{\Delta}_j^2},$$

where $\bar{\Delta}_j := \mu_j - S$.

For the other arms, the number of times they are uncertain is trivially bounded by the number of times they are played. However, we also show that insufficient arms are played almost exclusively during uncertain rounds.

Lemma B.3. *When Algorithm 1 is run with $C_2 \geq 6$, the probability that after T rounds insufficient arm i has been played in more than n rounds that are not uncertain is bounded as:*

$$\mathbb{P}\left[n_i(T+1) - n_{0,i}(T+1) > n\right] \leq \frac{2}{n^{C_2/2-1}}.$$

Then, the number of times an insufficient arm i is played can be decomposed as:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}\left[i(t)=i\right] \\
&\leq N_i + \sum_{t=1}^{T} \mathbb{P}\left[n_i(t) > N_i, i(t)=i\right] \\
&\leq N_i + \underbrace{\sum_{t=1}^{T} \mathbb{P}\left[i(t)=i, n_i(t) > N_i, u_{i^*} \leq \mu_{i^*} - \Delta_i/2\right]}_{\text{Cases where the best arm } i^* \text{ under-performs}} + \underbrace{\sum_{t=1}^{T} \mathbb{P}\left[i(t)=i, n_i(t) > N_i, u_i \geq \mu_i + \Delta_i/2\right]}_{\text{Cases where arm } i \text{ over-performs}},
\end{aligned} \tag{14}$$

where each term can be bounded according to the following to lemmas separately. Proofs of the lemmas are given in the following section.

Lemma B.4. *Consider any sub-Gaussian K-armed bandit problem. Using Algorithm 1 with $C_1 \geq 4$ and $C_2 \geq 6$, for any insufficient arm i and any $N_i \geq 1$, we have:*

$$\sum_{t=1}^{T} \mathbb{P}\left[i(t)=i, n_i(t) > N_i, u_{i^*} \leq \mu_{i^*} - \Delta_i/2\right] \leq \frac{21}{\Delta_i^2} + 16.$$

Lemma B.5. *Consider any sub-Gaussian two-armed bandit problem. Using Algorithm 1 with $C_1 \geq 4$ and $C_2 \geq 6$, if $N_2 \geq \frac{48}{\Delta_2^2} C_1 \log\left(3\frac{\log \frac{1}{\Delta_*}}{\Delta_*^2}\right) + \frac{144 C_1}{\Delta_2^2} \log\left(\frac{48 C_1}{\Delta_2^2}\right)$, then*

$$\sum_{t=1}^{T} \mathbb{P}\left[n_2(t) > N_2, i(t)=2, u_2(t) \geq \mu_2 + \Delta_2/2\right] \leq \frac{32}{\Delta_2^2} + \sum_{\kappa \geq 4} \kappa \Delta_*^{2\kappa-2} \log \frac{1}{\Delta_*}.$$

Combining these two results allows to prove Theorem 4.2.

Proof of Theorem 4.2. Decomposing the number of times suboptimal arm 2 is played using Equation (14) as well as Lemma B.4 and Lemma B.5, we deduce that:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}\left[i(t)=2\right] \\
&\leq \frac{48}{\Delta_2^2} C_1 \log\left(3\frac{\log \frac{1}{\Delta_*}}{\Delta_*^2}\right) + \frac{144 C_1}{\Delta_2^2} \log\left(\frac{48 C_1}{\Delta_2^2}\right) + \frac{21}{\Delta_2^2} + 16 + \frac{32}{\Delta_2^2} + \sum_{\kappa \geq 4} \kappa \Delta_*^{2\kappa-2} \log \frac{1}{\Delta_*} \\
&= O\left(\frac{1}{\Delta_2^2}\left(\log \frac{1}{\Delta_*} + \log \frac{1}{\Delta_2}\right) + \sum_{\kappa \geq 4} \kappa \Delta_*^{2\kappa-2} \log \frac{1}{\Delta_*}\right).
\end{aligned}$$

Plugging this bound in the definition of pseudo-regret, which coincides with the definition of the novel satisficing regret, finishes the proof. □

B.1 Proofs of the Lemmas

B.1.1 Properties of the Uncertain Rounds

Proof of Lemma B.2. For any sufficient arm j, we set $\bar{\Delta}_j := \mu_j - S$ and have:

$$\begin{aligned}
&\mathbb{P}\left[n_{0,j}(T+1) > n\right] \\
&= \mathbb{P}\left[\exists t : i(t) = j, n_{0,j}(t) = n, \hat{\mu}_j(t) < S + \sqrt{C_2 \frac{\log n_j(t)}{n_j(t)}}\right] \\
&\leq \sum_{n_j \geq n} \mathbb{P}\left[\exists t : i(t) = j, n_j(t) = n_j, \hat{\mu}_j(t) < S + \sqrt{C_2 \frac{\log n_j}{n_j}}\right] \\
&\leq \sum_{n_j \geq n} \exp\left(\frac{-n_j\left(\bar{\Delta}_j + \sqrt{C_2 \frac{\log n_j}{n_j}}\right)^2}{2}\right) \\
&\leq \sum_{n_j \geq n} \exp\left(-\frac{n_j \bar{\Delta}_j^2}{2}\right) \\
&\leq \frac{2\exp\left(-(n-1)\bar{\Delta}_j^2/2\right)}{\bar{\Delta}_j^2}.
\end{aligned}$$

The first step uses the definition of $n_{0,j}$. As we don't know how many times j is played, the second step takes a union bound on all possible values of $n_j(t)$, noting that $n_j(t) \geq n_{0,j}(t) = n$. In the third step, we use the fact that our rewards are sub-Gaussian. In the fourth step, we simplify using that $\exp(-x)$ is a decreasing function of x and that $(a+b)^2 \geq a^2 + b^2 \geq a^2$ for $a, b \geq 0$. The last step bounds the sum by an integral, as for any $c > 0$, we have $\sum_{n_i \geq n} \exp(-cn_i) \leq \lim_{x\to\infty} \int_{n-1}^{x} \exp(-cn_i) dn_i = \frac{\exp(-c(n-1))}{c}$.

□

Proof of Lemma B.3. For any insufficient arm i:

$$\begin{aligned}
&\mathbb{P}\left[n_i(T+1) - n_{0,i}(T+1) > n\right] \\
&\leq \sum_{n_i \geq n} \mathbb{P}\left[\exists t : i(t) = i, n_i(t) = n_i, \hat{\mu}_i(t) \geq S + \sqrt{C_2 \frac{\log n_i}{n_i}}\right] \\
&\leq \sum_{n_i \geq n} \mathbb{P}\left[\exists t : i(t) = i, n_i(t) = n_i, \hat{\mu}_i(t) \geq \mu_i + \sqrt{C_2 \frac{\log n_i}{n_i}}\right] \\
&\leq \sum_{n_i \geq n} \exp\left(\frac{-C_2 \log n_i}{2}\right) \\
&\leq \frac{2}{n^{C_2/2-1}}.
\end{aligned}$$

The first step counts the number of rounds where arm i is played and that round is not uncertain, and we use a union bound on the possible values of $n_i(t)$. In the second step, we upper bound by a looser condition as $S \geq \mu_i$. This allows the use of the properties of sub-Gaussian distributions. To finish the proof, we use that $C_2/2 \geq 2$ so $\frac{1}{n_i^{C_2/2}}$ can be integrated, and we upper bound the sum by an integral $\sum_{n_i \geq n} \frac{1}{n_i^{C_2/2}} \leq \lim_{x\to\infty} \int_{n-1}^{x} \frac{1}{n_i^{C_2/2}} = \frac{1}{(C_2/2-1)(n-1)^{C_2/2-1}} \leq \frac{1}{(n-1)^{C_2/2-1}} \leq \frac{2}{n^{C_2/2-1}}$. □

B.1.2 Handling the Insufficient Arms

Proof of Lemma B.4. For any insufficient arm i:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}\left[i(t)=i, n_i(t) > N_i, u_{i^*}(t) \leq \mu_{i^*} - \Delta_i/2\right] \\
&= \sum_{t=1}^{T} \mathbb{P}\left[i(t)=i, n_i(t) > N_i, \hat{\mu}_{i^*}(t) + \sqrt{C_1 \frac{\log n_{i^*}(t) + \log n_0(t)}{n_{i^*}(t)}} \leq \mu_{i^*} - \Delta_i/2\right] \\
&\leq \sum_{t=1}^{T} \mathbb{P}\left[i(t)=i, n_i(t) > N_i, \hat{\mu}_{i^*}(t) + \sqrt{C_1 \frac{\log n_{i^*}(t) + \log n_{0,i}(t)}{n_{i^*}(t)}} \leq \mu_{i^*} - \Delta_i/2\right] \\
&\leq \sum_{t=1}^{T} \mathbb{P}\left[i(t)=i, n_i(t) > N_i, \hat{\mu}_{i^*}(t) + \sqrt{C_1 \frac{\log n_{i^*}(t) + \log n_i(t)/2}{n_{i^*}(t)}} \leq \mu_{i^*} - \Delta_i/2\right] \\
&\quad + \sum_{t=1}^{T} \mathbb{P}\left[i(t)=i, n_i(t) > N_i, n_{0,i}(t) < n_i(t)/2\right]. \qquad (15)
\end{aligned}$$

The first step replaces $u_{i^*}(t)$ by its definition. The second step lower bounds $n_0(t) \geq n_{0,i}(t)$. In the third step, we decompose that probability by taking the cases where $n_{0,i}(t) \geq n_i(t)/2$ or $n_{0,i}(t) < n_i(t)/2$.

For the first term of Equation (15), we have:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}\left[i(t)=i, n_i(t) > N_i, \hat{\mu}_{i^*}(t) + \sqrt{C_1 \frac{\log n_{i^*}(t) + \log n_i(t)/2}{n_{i^*}(t)}} \leq \mu_{i^*} - \Delta_i/2\right] \\
&\leq \sum_{n_i=N_i+1}^{T} \sum_{n_{i^*}=1}^{T} \mathbb{P}\left[\exists t : \hat{\mu}_{i^*}(t) = \hat{\mu}_{i^*,n_i^*}, \hat{\mu}_{i^*,n_i^*} + \sqrt{C_1 \frac{\log n_{i^*} + \log n_i/2}{n_{i^*}}} \leq \mu_{i^*} - \Delta_i/2\right] \\
&\leq \sum_{n_i=N_i+1}^{T} \sum_{n_{i^*}=1}^{T} \mathbb{P}\left[\exists t : \hat{\mu}_{i^*}(t) = \hat{\mu}_{i^*,n_i^*}, \hat{\mu}_{i^*,n_i^*} \leq \mu_{i^*} - \Delta_i/2 - \sqrt{C_1 \frac{\log n_{i^*} + \log n_i/2}{n_{i^*}}}\right] \\
&\leq \sum_{n_i=N_i+1}^{T} \sum_{n_{i^*}=1}^{T} \exp\left(-\frac{n_{i^*}}{2}\left(\Delta_i/2 + \sqrt{C_1 \frac{\log n_{i^*} + \log n_i/2}{n_{i^*}}}\right)^2\right) \\
&\leq \sum_{n_i=N_i+1}^{T} \sum_{n_{i^*}=1}^{T} \exp\left(-\frac{n_{i^*}}{8}\Delta_i^2 - \frac{C_1}{2}\log n_{i^*} - \frac{C_1}{2}\log\frac{n_i}{2}\right) \\
&\leq \sum_{n_{i^*}=1}^{T} \exp\left(-\frac{n_{i^*}}{8}\Delta_i^2\right) \sum_{n_i=N_i+1}^{T} \exp\left(-\frac{C_1}{2}\log\frac{n_i}{2}\right) \\
&\leq \frac{21}{\Delta_i^2},
\end{aligned}$$

where we first take an upper bound by taking all the possible values n_{i^*} and n_i instead of summing solely on t. Removing the unnecessary dependencies, we can apply our sub-Gaussian assumption as defined in Equation (1) and simplify using that $\exp(-x)$ is a decreasing function of x and that $(a+b)^2 \geq a^2 + b^2$ for $a, b \geq 0$. Using Proposition A.1, we can decompose the result into a product of two convergent sums. The first sum is bounded by $\frac{8}{\Delta_i^2}$ and the second can be loosely bounded as $\sum_{n_i=N_i+1}^{T} \exp\left(-\frac{C_1}{2}\log\frac{n_i}{2}\right) = \sum_{n_i=N_i+1}^{T} \left(\frac{2}{n_i}\right)^{C_1/2} \leq \sum_{n_i=2}^{T} \left(\frac{2}{n_i}\right)^2 \leq 2.6$ when $C_1 \geq 4$ (note that we are summing from $n_i = 2$, so each term $0 < \frac{2}{n_i} \leq 1$ and thus we take an upper bound by lower bounding the exponent).

For the second sum in Equation (15):

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}\left[i(t)=i, n_i(t) > N_i, n_{0,i}(t) < n_i(t)/2\right] \\
&\leq \sum_{t=1}^{T} \mathbb{P}\left[i(t)=i, n_i(t) > N_i, n_{0,i}(t) < n_i(t) - n_i(t)/2\right] \\
&\leq \sum_{n_i=N_i+1}^{T} \mathbb{P}\left[n_i(T+1) - n_{0,i}(T+1) > n_i/2\right] \\
&\leq \sum_{n_i=N_i+1}^{T} \frac{2}{(n_i/2)^{C_2/2-1}} \\
&\leq 16,
\end{aligned}$$

where we rewrite $n_i(t)/2 = n_i(t) - n_i(t)/2$. Furthermore, $n_i(t) - n_{0,i}(t)$ is an increasing function of t, so we can rearrange the terms and $n_i(t) - n_{0,i}(t)$ is upper bounded by $n_i(T+1) - n_{0,i}(T+1)$. Then, we take an upper bound by summing on all possible values of n_i rather than on the possible values of t. We can apply Lemma B.3 with $n = n_i/2$. For any $C_2 \geq 6$, we have $\sum_{n_i=N_i+1}^{T} \frac{2}{(n_i/2)^{C_2/2-1}} \leq \sum_{n_i=N_i+1}^{T} \frac{2}{(n_i/2)^2} = 8\sum_{n_i=1}^{T} \frac{1}{n_i^2} \leq 16$.

We conclude that for any $N_i \geq 1$:

$$\sum_{t=1}^{T} \mathbb{P}\left[i(t)=i, n_i(t) > N_i, u_{i^*} \leq \mu_{i^*} - \Delta_i/2\right] \leq \frac{21}{\Delta_i^2} + 16.$$

□

Proof of Lemma B.5. Without loss of generality, we assume that arm 1 is the optimal arm and arm 2 is sub-optimal. Then, we have:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}\left[n_2(t) > N_2, i(t)=2, u_2(t) \geq \mu_2 + \Delta_2/2\right] \\
&= \sum_{t=1}^{T} \sum_{n_2>N_2} \mathbb{P}\left[n_2(t)=n_2, i(t)=2, \hat{\mu}_2(t) + \sqrt{C_1 \frac{\log n_2(t) + \log n_0(t)}{n_2(t)}} \geq \mu_2 + \Delta_2/2\right] \\
&\leq \sum_{t=1}^{T} \sum_{n_2>N_2} \mathbb{P}\left[n_2(t)=n_2, i(t)=2, \hat{\mu}_2(t) + \sqrt{C_1 \frac{2\log n_2(t) + \log n_{0,1}(T+1)}{n_2(t)}} \geq \mu_2 + \Delta_2/2\right] \\
&\leq \sum_{n_2>N_2} \mathbb{P}\left[\hat{\mu}_{2,n_2} + \sqrt{C_1 \frac{2\log n_2 + \log n_{0,1}(T+1)}{n_2}} \geq \mu_2 + \Delta_2/2\right] \\
&\leq \sum_{n_2>N_2} \mathbb{P}\left[\hat{\mu}_{2,n_2} + \sqrt{C_1 \frac{2\log n_2 + \log(3N_{0,1})}{n_2}} \geq \mu_2 + \Delta_2/2\right] \\
&+ \sum_{n_2>N_2} \sum_{\kappa\geq 4} \mathbb{P}\left[(\kappa-1)N_{0,1} < n_{0,1}(T+1) \leq \kappa N_{0,1}, \hat{\mu}_{2,n_2} + \sqrt{C_1 \frac{2\log n_2 + \log(\kappa N_{0,1})}{n_2}} \geq \mu_2 + \Delta_2/2\right]
\end{aligned} \tag{16}$$

$$\leq \sum_{n_2>N_2} \mathbb{P}\left[\hat{\mu}_{2,n_2} + \sqrt{C_1 \frac{2\log n_2 + \log(3N_{0,1})}{n_2}} \geq \mu_2 + \Delta_2/2\right]$$
$$+ \sum_{n_2>N_2} \sum_{\kappa:n_2\geq\kappa N_{0,1}} \mathbb{P}\left[(\kappa-1)N_{0,1} < n_{0,1}(T+1) \leq \kappa N_{0,1}, \hat{\mu}_{2,n_2} + \sqrt{C_1 \frac{2\log n_2 + \log n_2}{n_2}} \geq \mu_2 + \Delta_2/2\right]$$
$$+ \sum_{n_2>N_2} \sum_{\kappa:n_2<\kappa N_{0,1}} \mathbb{P}\left[(\kappa-1)N_{0,1} < n_{0,1}(T+1) \leq \kappa N_{0,1}\right], \tag{17}$$

where the first step replaces $u_i(t)$ by its definition and the second step replaces $n_0(t)$. Then, we re-index the sum in terms of n_i rather than t. In Equation (16), we decompose the $n_{0,1}(T)$ according to its value compared to some constant $N_{0,1}$. This sum is then decomposed again depending on whether κ fulfills $n_2 \geq \kappa N_{0,1}$ or not.

We now bound each of the three terms in Equation (17) separately.

Focusing on the first sum and following the same approach as before, we re-index the sum by only considering the values of $n_i(t)$ rather than all t.

$$\sum_{n_2>N_2} \mathbb{P}\left[\hat{\mu}_{2,n_2} + \sqrt{C_1 \frac{2\log n_2 + \log(3N_{0,1})}{n_2}} \geq \mu_2 + \Delta_2/2\right]$$
$$\leq \sum_{n_2>N_2} \exp\{-n_2\Delta_2^2/4 + 2C_1 \log n_2 + C_1 \log(2N_{0,1})\}$$
$$\leq 2 \sum_{n_2>N_2} \exp\{-n_2\Delta_2^2/8\}$$
$$\leq \frac{16}{\Delta_2^2}$$

when N_2 is large enough. For this condition to hold, we need

$$\sqrt{C_1 \frac{2\log n_2 + \log(3N_{0,1})}{n_2}} \geq \Delta_2/4. \tag{18}$$

Choosing $N_{0,1} = \frac{\log \frac{1}{\Delta_*}}{\Delta_*^2}$, we note that the LHS of Equation (18) is a decreasing function of n_2, and so it is sufficient to find an initial value of n_2 that fulfills this condition. Reorganizing the terms gives:

$$n_2 \geq \text{argmin}_{n_2}\left(\sqrt{C_1 \frac{2\log n_2 + \log(3N_{0,1})}{n_2}} \geq \Delta_2/4\right)$$
$$\Leftrightarrow\ n_2 \geq \frac{16}{\Delta_2^2} C_1 \left(2\log n_2 + \log\left(3N_{0,1}\right)\right) \tag{19}$$

To upper bound this quantity, we consider two cases: either $n_2 \leq 3N_{0,1}$ or $n_2 > 3N_{0,1}$. When $n_2 > 3N_{0,1}$ then Equation (19) implies:

$$n_2 \geq \frac{48}{\Delta_2^2} C_1 \log\left(3N_{0,1}\right) \geq \frac{48}{\Delta_2^2} C_1 \log\left(3\frac{\log \frac{1}{\Delta_*}}{\Delta_*^2}\right).$$

If $n_2 \leq 3N_{0,1}$ then we obtain from Equation (19):

$$n_2 \geq \frac{48}{\Delta_2^2} C_1 \left(\log n_2\right). \tag{20}$$

This expression still needs to be solved for n_2. To do so, we use the same approach as Rouyer et al. (2022) and use W_{-1} to denote the product log function, as defined by Chatzigeorgiou (2013, Theorem 1). We have that for any constant $c \geq e$, the solution of $n = c \log n$ is $n = -cW_{-1}\left(-\frac{1}{c}\right)$. Then,

$$-cW_{-1}\left(-\frac{1}{c}\right) = -cW_{-1}\left(-\exp\left(-\log\left(\frac{c}{e}\right)\right) - 1\right) \leq c\left(1 + \sqrt{2\log\left(\frac{c}{e}\right)} + \log\left(\frac{c}{e}\right)\right).$$

We can use this upper bound with $c = \frac{48C_1}{\Delta_2^2}$, and assuming that $c \geq e^2$, we have

$$\left(1 + \sqrt{2\log\left(\frac{c}{e}\right)} + \log\left(\frac{c}{e}\right)\right) \leq \left(1 + \sqrt{2\log c} + \log c\right) \leq 3\log c,$$

from which we deduce that:

$$n_2 \geq \frac{144C_1}{\Delta_2^2}\log\left(\frac{48C_1}{\Delta_2^2}\right). \tag{21}$$

Combining Equation (20) and Equation (21) allows to pick

$$N_2 = \frac{48}{\Delta_2^2}C_1\log\left(3\frac{\log\frac{1}{\Delta_*}}{\Delta_*^2}\right) + \frac{144C_1}{\Delta_2^2}\log\left(\frac{48C_1}{\Delta_2^2}\right), \tag{22}$$

which ensures that the terms in Equation (17) can be bounded properly.

Moving on to the second sum in Equation (17), this can be upper bounded by

$$\begin{aligned}
&\sum_{n_2 > N_2} \mathbb{P}\left[\hat{\mu}_{2,n_2} + \sqrt{C_1\frac{2\log n_2 + \log n_2}{n_2}} \geq \mu_2 + \Delta_2/2\right] \\
&\leq \sum_{n_2 > N_2} \exp\{-n_2\Delta_2^2/4 + 3C_1\log n_2\} \\
&\leq 2\sum_{n_2 > N_2} \exp\{-n_2\Delta_2^2/8\} \\
&\leq \frac{16}{\Delta_2^2},
\end{aligned}$$

following the same approach as before.

Finally, for the last term in Equation (17) we have

$$\begin{aligned}
&\sum_{\kappa \geq 4}\sum_{n_2 : n_2 < \kappa N_{0,1}} \mathbb{P}\left[(\kappa-1)N_{0,1} < n_{0,1}(T+1) \leq \kappa N_{0,1}\right] \\
&\leq \sum_{\kappa \geq 4} \kappa N_{0,1}\exp\{-(\kappa-1)N_{0,1}\Delta_*^2\}/\Delta_*^2 \\
&\leq \sum_{\kappa \geq 4} \kappa N_{0,1}\Delta_*^{2\kappa-4} \\
&\leq \sum_{\kappa \geq 4} \kappa\frac{\log\frac{1}{\Delta_*}}{\Delta_*^2}\Delta_*^{2\kappa-4},
\end{aligned}$$

which converges when $\Delta_* < 1$.

To conclude, we proved that if $N_2 \geq \frac{48}{\Delta_2^2}C_1\log\left(3\frac{\log\frac{1}{\Delta_*}}{\Delta_*^2}\right) + \frac{144C_1}{\Delta_2^2}\log\left(\frac{48C_1}{\Delta_2^2}\right)$, then

$$\sum_{t=1}^{T}\mathbb{P}\left[n_2(t) > N_2, i(t) = 2, u_2(t) \geq \mu_2 + \Delta_2/2\right] \leq \frac{32}{\Delta_2^2} + \sum_{\kappa \geq 4}\kappa\Delta_*^{2\kappa-4}\frac{\log\frac{1}{\Delta_*}}{\Delta_*^2}.$$

□