Regret Bounds for Satisficing in Multi-Armed Bandit Problems

Thomas Michel thomas.michel1@ens-paris-saclay.fr
Université Paris-Saclay, ENS Paris-Saclay
France

Hossein Hajiabolhassan hossein.hajiabolhassan@unileoben.ac.at
Lehrstuhl für Informationstechnologie
Montanuniversität Leoben
Austria

Ronald Ortner rortner@unileoben.ac.at
Lehrstuhl für Informationstechnologie
Montanuniversität Leoben
Austria

Abstract

This paper considers the objective of *satisficing* in multi-armed bandit problems. Instead of aiming to find an optimal arm, the learner is content with an arm whose reward is above a given satisfaction level. We provide algorithms and analysis for the realizable case when such a satisficing arm exists as well as for the general case when this may not be the case. Introducing the notion of *satisficing regret*, our main result shows that in the general case it is possible to obtain constant satisficing regret when there is a satisficing arm (thereby correcting a contrary claim in literature), while standard logarithmic regret bounds can be re-established otherwise. Experiments illustrate that our algorithm is not only superior to standard algorithms in the satisficing setting, but also works well in the classic bandit setting.

Keywords: multi-armed bandit, satisficing, regret

1. Introduction

One of the reasons why reinforcement learning (RL) is in general difficult is that finding an *optimal* policy in general requires a lot of exploration. In practice however, we are often happy to perform a task just good enough. For example, when driving to work we will be content with a strategy that will let us arrive just in time, while the computation of a policy that is 'optimal' in some sense (e.g., along the shortest route, or as fast as possible) may be prohibitive. Accordingly, it is to be expected that when considering a *satisficing* objective aiming to find a solution that is above a certain satisfaction level it is possible to learn a respective policy much faster.

While there are some connections to multi-criterion RL (Roijers et al., 2013), there is hardly any literature on satisficing in RL, with a few exceptions for the multi-armed bandit (MAB) setting. Kohno and Takahashi (2017) and Tamatsukuri and Takahashi (2019) propose simple index policies, which are experimentally evaluated. Tamatsukuri and Takahashi (2019) also show that the suggested algorithm converges to a satisficing arm and that the regret is finite if the satisfaction level is chosen to be between the reward of the best and the second-best arm.

Reverdy et al. (2017) consider a more general Bayesian setting, which also considers the learner's belief that some arm is satisficing. The notion of *expected satisficing regret* is introduced that measures the loss over all

steps where a non-satisficing arm is chosen and the learner's degree of belief in the chosen arm was below some level $\delta \in [0, 1]$. For $\delta = 0$ this coincides with our notion of *satisficing regret* introduced below. Reverdy et al. (2017) present various bounds on the expected satisficing regret, including lower bounds as well as upper bounds for problems with Gaussian reward distributions when using adaptations of the UCL algorithm (Reverdy et al., 2014). The given bounds for the case $\delta = 0$ that correspond to our setting will be discussed in Section 3 below.

Merlis and Mannor (2021) consider a related notion of so-called *lenient regret* that considers the loss with respect to $\mu_* - \varepsilon$ for a parameter $\varepsilon > 0$ that specifies the allowed deviation from the optimal mean reward μ_*. Asymptotic upper bounds on the lenient regret are shown for a version of Thompson sampling that match a given lower bound. When $\mu_* > 1 - \varepsilon$ the lenient regret turns out to be constant.

Russo and Roy (2018) consider satisficing in a setting with discounted rewards and provide respective bounds on the expected discounted regret for a satisficing variant of Thompson sampling (Thompson, 1933).

Also related to our paper, Kano et al. (2019) consider the problem of identifying *all* arms above a given satisfaction level and derive sample complexity bounds for the pure-exploration setting with fixed confidence. Related sample complexity bounds can be found in (Mason et al., 2020) for identification of all ε-good arms. Closer to our setting is the problem of identifying an arbitrary arm among the top m arms, for which sample complexity bounds are derived by Chaudhuri and Kalyanakrishnan (2017). A follow-up paper (Chaudhuri and Kalyanakrishnan, 2019) considers the sample complexity of the more general problem of identification of any k of the best m arms. None of these latter investigations however considers the online learning setting with regret as performance measure as we do. Note that an algorithm for pure exploration after any number of steps with high probability will identify an optimal or at least a satisficing arm. However, subsequent exploitation will always give linear regret due to the small but positive error probability so that a simple approach of first exploring and then exploiting does not work well in general.

Investigating also the MAB setting, in this paper we first introduce the notion of *satisficing regret* that measures the loss with respect to a given satisfaction level S. We first consider the realizable case, where this level can be satisfied. In this setting quite a simple algorithm can be shown to have constant satisficing regret (i.e., no dependence on the horizon T). For the general setting we provide an algorithm that is able to extend on this result, giving constant satisficing regret in the realizable case, while obtaining logarithmic bounds on the ordinary regret with respect to the optimal arm as for classic MAB algorithms such as UCB1 (Auer et al., 2002). Experiments not only confirm our theoretical findings but also show that our algorithm is competitive even in the standard setting.

2. Setting

We consider the standard multi-armed bandit (MAB) setting with a set of K arms given, in the following denoted as $[\![1, K]\!] := \{1, 2, \ldots, K\}$. In discrete time steps $t = 1, 2, \ldots$ the learner picks an arm $A_t = i$ from $[\![1, K]\!]$ and observes a random reward r_t drawn from a fixed reward distribution specific to the chosen arm i with mean μ_i. In the following we assume that the reward distributions for each arm are sub-Gaussian. This is e.g. guaranteed when the reward distributions are bounded, which is a common assumption in the bandit setting.

The usual performance measure for a learning algorithm in the MAB setting is the *(pseudo-)regret* after T steps, defined as

$$R_T := \sum_{t=1}^{T} \big(\mu_* - \mu_{A_t}\big),$$

where $\mu_* := \max_i \mu_i$ is the maximal mean reward of all arms.

In the satisficing setting however, we only care about whether an arm with mean reward $\geqslant S$ is chosen, where S is the level of satisfaction we aim at. Accordingly, we modify the classic notion of regret and consider what we call the *satisficing (pseudo-)regret* with respect to S (short S*-regret*) defined as

$$R_T^S := \sum_{t=1}^{T} \max\{S - \mu_{A_t}, 0\}.$$

This definition reflects that we are happy with any arm having mean reward $\geqslant S$ and that there is no benefit in overfulfilling the given satisfaction level S. Note that the S-regret will be linear in T whenever there is no satisficing arm with mean reward $\geqslant S$, that is, if $\mu_* < S$. As already mentioned, a more general notion of satisficing regret in a Bayesian setting that also considers the learner's degree of belief and coincides with S-regret in our particular setting has been suggested by Reverdy et al. (2017).

3. The Realizable Case

We start with the *realizable case* when $\mu_* > S$. The main goal of this section is to show that suitable algorithms will have just constant S-regret in this case. Note that this does not hold for standard algorithms like UCB1 (Auer et al., 2002). Lower bounds show that these algorithms will choose a suboptimal arm i for $\Omega\Big(\frac{logT}{(\mu^*-\mu_i)^2}\Big)$ times. This of course also holds for any arm below the satisfaction level S giving a contribution to the overall S-regret of $\Omega\Big(\frac{(S-\mu_i)\log T}{(\mu^*-\mu_i)^2}\Big)$.

3.1 Simple Algorithm

We start with a simple algorithm shown as Algorithm 1. It plays the empirical best arm so far if its empirical mean reward is $\geqslant S$ and explores uniformly at random otherwise. In the following, the empirical reward for arm i available at step t (i.e., *before* choosing the arm A_t) is denoted by $\hat{\mu}_i(t)$.

Algorithm 1

Require: K, S
1: Play each arm once, i.e., for time steps $t = 1, \ldots, K$ play arm $A_t = t$.
2: **for** time steps $t = K + 1, \ldots$ **do**
3: **if** $\exists i\, \hat{\mu}_i(t) \geqslant S$ **then**
4: Play $A_t \leftarrow \operatorname{argmax}_{i \in [\![1,K]\!]} \hat{\mu}_i(t)$.
5: **else**
6: Choose A_t uniformly at random from $[\![1, K]\!]$.
7: **end if**
8: **end for**

Analogously to the ordinary MAB setting where the gaps $\Delta_i := \mu_* - \mu_i$ to the optimal arm appear in bounds on the (classic) regret, when satisficing the gaps $\Delta_i^S = S - \mu_i$ for non-satisficing arms are important parameters describing the difficulty of the problem. Indeed, one can show the following bound on the S-regret.

Theorem 1. *If $S < \mu_*$ then Algorithm 1 satisfies for all $T \geqslant 1$*

$$R_T^S \leqslant \sum_{i:\Delta_i^S > 0} \Big(\Delta_i^S + \frac{2}{\Delta_i^S} + \frac{2\Delta_i^S}{|\Delta_*^S|^2}\Big).$$

For the proof we shall need the following result that follows by our assumption of sub-Gaussianity and a Chernoff bound.

Lemma 2. *Let $\hat{\mu}_{i,n}$ be an empirical estimate for μ_i computed from n samples. Then for all $\varepsilon > 0$ and each $i \in [\![1, K]\!]$,*

$$\mathbb{P}(\hat{\mu}_{i,n} \geqslant \mu_i + \varepsilon) \leqslant \exp(-\tfrac{n\varepsilon^2}{2}),$$
$$\mathbb{P}(\hat{\mu}_{i,n} \leqslant \mu_i - \varepsilon) \leqslant \exp(-\tfrac{n\varepsilon^2}{2}).$$

Proof of Theorem 1. Let i be the index of a non-satisficing arm. In the following we decompose the event that arm i is chosen at some step t. To do that we introduce the event $Z_t := \{\forall j \in [\![1, K]\!], \hat{\mu}_i(t) < S\}$ that all arms have empirical estimates below S, when the algorithm chooses an arm randomly according to line 6 of the algorithm. Then we have

$$\{A_t = i\} \subset \{t = i\} \cup \{A_t = i, Z_t^c\} \cup \{A_t = i, Z_t\}. \tag{1}$$

For the first two events we have

$$\sum_{t=1}^{T} \mathbb{P}(t = i) \leqslant 1 \tag{2}$$

and

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{P}(A_t = i, Z_t^c) &\leqslant \sum_{t=1}^{T} \mathbb{P}\big(A_t = i, \hat{\mu}_i(t) \geqslant S\big) \\
&= \sum_{t=1}^{T} \mathbb{P}\big(A_t = i, \hat{\mu}_i(t) \geqslant \mu_i + \Delta_i^S\big) \leqslant \sum_{n=1}^{T} \mathbb{P}(\hat{\mu}_{i,n} \geqslant \mu_i + \Delta_i^S) \\
&\leqslant \sum_{n=1}^{T} \exp\big(-\tfrac{n(\Delta_i^S)^2}{2}\big) \leqslant \frac{e^{-\frac{(\Delta_i^S)^2}{2}}}{1 - e^{-\frac{(\Delta_i^S)^2}{2}}} \leqslant \frac{2}{(\Delta_i^S)^2}.
\end{aligned} \tag{3}$$

Rewriting the probability of the third event in (2), using $*$ to refer to an arbitrary optimal arm, we obtain

$$\begin{aligned}
\mathbb{P}(A_t = i, Z_t) &= \mathbb{P}(A_t = i | Z_t)\, \mathbb{P}(Z_t) = \tfrac{1}{K} \cdot \mathbb{P}(Z_t) \\
&= \mathbb{P}(A_t = * | Z_t)\, \mathbb{P}(Z_t) = \mathbb{P}(A_t = *, Z_t).
\end{aligned}$$

Now summing over the time steps up to T yields

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{P}(A_t = i, Z_t) &= \sum_{t=1}^{T} \mathbb{P}(A_t = *, Z_t) \leqslant \sum_{t=1}^{T} \mathbb{P}\big(A_t = *, \hat{\mu}_*(t) \leqslant S\big) \\
&= \mathbb{E}\Big(\sum_{t=1}^{T} \mathbb{1}\big\{A_t = *, \hat{\mu}_*(t) \leqslant S\big\}\Big) \leqslant \mathbb{E}\Big(\sum_{n=1}^{T} \mathbb{1}\{\hat{\mu}_{*,n} \leqslant S\}\Big) \\
&= \sum_{n=1}^{T} \mathbb{P}(\hat{\mu}_{*,n} \leqslant S) = \sum_{n=1}^{T} \mathbb{P}\big(\hat{\mu}_{*,n} \leqslant \mu_* - |\Delta_*^S|\big) \\
&\leqslant \sum_{n=1}^{T} \exp\big(-\tfrac{n|\Delta_*^S|^2}{2}\big) \leqslant \frac{2}{|\Delta_*^S|^2}.
\end{aligned} \tag{4}$$

Finally writing

$$n_i(T) = \sum_{t=1}^{T} \mathbb{I}\{A_t = i\}$$

for the number of times arm i was pulled up to step T, we can combine (1)–(4) to obtain

$$\begin{aligned} R_T^S &= \sum_{i:\Delta_i^S>0} \Delta_i^S \, \mathbb{E}(n_i(T)) \;=\; \sum_{i:\Delta_i^S>0} \Delta_i^S \sum_{t=1}^{T} \mathbb{P}(A_t = i) \\ &\leqslant \sum_{i:\Delta_i^S>0} \Delta_i^S \left(1 + \frac{2}{(\Delta_i^S)^2} + \frac{2}{|\Delta_*^S|^2}\right) \leqslant \sum_{i:\Delta_i^S>0} \left(\Delta_i^S + \frac{2}{\Delta_i^S} + \frac{2\Delta_i^S}{|\Delta_*^S|^2}\right). \end{aligned}$$ □

The algorithm as well as the analysis are adaptations from Bubeck et al. (2013) where ordinary regret bounds for the MAB setting are considered under the assumption that the learner knows the value of μ_* as well as (a bound on) the gap Δ between the optimal and the best suboptimal arm.[1] The crucial insight is that what is actually needed in order to apply algorithm and analysis of Bubeck et al. (2013) is to have a reference value μ that separates the optimal from suboptimal arms, that is, $\mu_* > \mu > \mu_i$ for all suboptimal arms i. In our case this reference value is given by the satisfaction level S, which in the realizable case separates the good arms from the bad ones. Note that for this we need to have $S < \mu_*$, so that we do not get constant regret when $S = \mu_*$.

Remark. (i) Concerning lower bounds, choosing a scenario with a single optimal arm with $\mu_* = S + \Delta$ and suboptimal arms i with mean reward $\mu_i = S - \Delta$ so that $\Delta_*^S = \Delta_i^S$, Theorem 5 of Bubeck et al. (2013) shows that the satisfying regret for any algorithm is of order $1/\Delta$, which coincides with the bound of Theorem 1.

(ii) The constant regret bound of Theorem 1 not only improves over the logarithmic bounds given by Reverdy et al. (2017) for a variant of the UCL algorithm (Reverdy et al., 2014) that picks an arbitrary arm with an UCL-index above S (instead of an arm with maximal index). Our bound also is not consistent with a claimed lower bound that is also logarithmic in the horizon (not mentioned in the corrections of Reverdy et al., 2021). This bound is obtained by application of a lower bound for the *multiple play* setting (Anantharam et al., 1987), where at each step m arms are chosen by the learner, who hence has to identify the m best arms. The given proof chooses m to be all arms above the given satisfaction level S. However, the lower bound is obviously not directly applicable to the satisficing setting: not *all* arms above the satisfaction level have to be found, but a single one is sufficient.

Bubeck et al. (2013) also provide another algorithm with a more refined approach for exploration, using a potential function instead of a uniform probability distribution over the arms. An adaptation to the satisficing setting is given in Appendix A, which also includes a regret analysis for the respective Algorithm 2.

4. The General Case

Now let us consider the general case where it is not guaranteed that the chosen satisfaction level S is realizable, that is, it may happen that $S > \mu_*$. Then unlike in the realizable case the satisfaction level S does not give the learner any useful information so that we cannot hope to perform better than in an ordinary MAB setting. Obviously the S-regret will be linear, but we aim at getting bounds on the (classic) regret. On the other hand,

1. As has been shown in the meantime knowledge of μ_* is sufficient for obtaining constant bounds on the regret (Garivier et al., 2019).

if there is at least one arm above the satisfaction level S, we would like to re-establish constant bounds on the S-regret as in the realizable case.

For the general setting we propose Algorithm 3. It uses a more refined approach for exploitation. Instead of just deciding based on the empirical estimate $\hat{\mu}_i$ of each arm i, it also considers for each arm a confidence interval defined by the two values (the first one being similar to the classical value suggested for the UCB1 algorithm of Auer et al., 2002)

$$\mathrm{UCB}_i(t) := \hat{\mu}_i(t) + \beta_i(t), \quad \text{and} \quad \mathrm{LCB}_i(t) := \hat{\mu}_i(t) - \beta_i(t), \tag{5}$$
$$\text{where } \beta_i(t) = \sqrt{\tfrac{2\log(f(t))}{n_i(t-1)}} \text{ with } f(t) = 1 + t\log^2(t).$$

If there is an arm with empirical mean above S, the algorithm chooses the arm for which the largest share of this confidence interval is above the satisfaction level S (cf. line 4 of the algorithm). Otherwise, if there is at least one arm with UCB-value above S such an arm is chosen uniformly at random. If all arms have UCB-value below S, the algorithm chooses an arm according to UCB1, that is, an arm i maximizing UCB_i.

Algorithm 3

Require: K, S
1: Play each arm once, i.e., for time steps $t = 1, \dots, K$ play arm $A_t = t$.
2: **for** time steps $t = K+1, \dots$ **do**
3: **if** $\exists i\, \mu_i(t) \geqslant S$ **then**
4: Choose $A_t \in \operatorname{argmax}_{i \in [\![1,K]\!]} \left\{ \frac{\mathrm{UCB}_i(t) - \max\{S, \mathrm{LCB}_i(t)\}}{\beta_i(t)} \right\}$.
5: **else if** $\exists i\ \mathrm{UCB}_i(t) \geqslant S$ **then**
6: Choose A_t uniformly at random from $\{i \mid \mathrm{UCB}_i(t) \geqslant S\}$.
7: **else**
8: Play arm $A_t \in \operatorname{argmax}_{i \in [\![1,K]\!]} \mathrm{UCB}_i(t)$.
9: **end if**
10: **end for**

The following two theorems show that Algorithm 3 is able to achieve constant S-regret if $\mu_* > S$, while the regret is bounded as for UCB1 otherwise (Auer et al., 2002).

Theorem 3. *If $\mu_* > S$ then Algorithm 3 satisfies for all $T \geqslant 1$,*

$$R_T^S \leqslant \sum_{i:\Delta_i^S > 0} \left(\Delta_i^S + \frac{2}{\Delta_i^S} + \frac{7\Delta_i^S}{|\Delta_*^S|^2} \right).$$

Proof. As before we write the S-regret as

$$R_T^S = \sum_{i:\Delta_i^S > 0}^{k} \mathbb{E}(n_i(T))\, \Delta_i^S$$

and proceed bounding $\mathbb{E}(n_i(T)) = \sum_{t=1}^{T} \mathbb{P}(A_t = i)$ for all non-satisficing arms i. Thus let i be the index of a non-satisficing arm. Let $Z_t := \{\forall j \in [\![1, K]\!], \hat{\mu}_j(t) < S\}$ be again the event that all arms have empirical

values below the satisfaction level. Then we can decompose the event $\{A_t = i\}$ as

$$\begin{aligned}\{A_t = i\} \subset \{t = i\} &\cup \big\{A_t = i, \hat{\mu}_i(t) \geqslant S, t > K\big\} \\ &\cup \big\{A_t = i, \mathrm{UCB}_i(t) \geqslant S, \mathrm{UCB}_*(t) \geqslant S, t > K, Z_t\big\} \\ &\cup \big\{A_t = i, \mathrm{UCB}_*(t) < S, t > K, Z_t\big\}. \end{aligned} \tag{6}$$

For the first two events we have

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{P}(t = i) \leqslant 1 \quad \text{and} \quad \sum_{t=1}^{T} \mathbb{P}\big(A_t = i, \hat{\mu}_i(t) \geqslant S, t > K\big) &\leqslant \sum_{t=1}^{T} \mathbb{P}\big(A_t = i, \hat{\mu}_i(t) \geqslant \mu_i + \Delta_i^S\big) \\ &\leqslant \sum_{n=1}^{T} \mathbb{P}(\hat{\mu}_{i,n} \geqslant \mu_i + \Delta_i^S) \leqslant \sum_{n=1}^{T} \exp\big(-\tfrac{n(\Delta_i^S)^2}{2}\big) \\ &\leqslant \frac{e^{-\frac{(\Delta_i^S)^2}{2}}}{1 - e^{-\frac{(\Delta_i^S)^2}{2}}} \leqslant \frac{2}{(\Delta_i^S)^2}\,. \end{aligned} \tag{7}$$

For the probability of the third event we have

$$\begin{aligned}&\sum_{t=1}^{T} \big\{A_t = i, \mathrm{UCB}_i(t) \geqslant S, \mathrm{UCB}_*(t) \geqslant S, t > K, Z_t\big\} \\ &\quad= \sum_{t=1}^{T} \big\{A_t = *, \mathrm{UCB}_i(t) \geqslant S, \mathrm{UCB}_*(t) \geqslant S, t > K, Z_t\big\} \\ &\quad\leqslant \sum_{t=1}^{T} \mathbb{P}(A_t = *, Z_t) \leqslant \sum_{t=1}^{T} \mathbb{P}\big(A_t = *, \hat{\mu}_*(t) \leqslant S\big) \\ &\quad= \mathbb{E}\Big(\sum_{t=1}^{T} \mathbb{1}\big\{A_t = *, \hat{\mu}_*(t) \leqslant S\big\}\Big) \leqslant \mathbb{E}\Big(\sum_{n=1}^{T} \mathbb{1}\{\hat{\mu}_{*,n} \leqslant S\}\Big) \\ &\quad= \sum_{n=1}^{T} \mathbb{P}(\hat{\mu}_{*,n} \leqslant S) = \sum_{n=1}^{T} \mathbb{P}\big(\hat{\mu}_{*,n} \leqslant \mu_* - |\Delta_*^S|\big) \\ &\quad\leqslant \sum_{n=1}^{T} \exp\big(-\tfrac{n|\Delta_*^S|^2}{2}\big) \leqslant \frac{2}{|\Delta_*^S|^2}. \end{aligned} \tag{8}$$

Finally, the probability of the last event of (6) is upper bounded by

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{P}\big(\mathrm{UCB}_*(t) < S\big) &= \sum_{t=1}^{T} \mathbb{P}\big(\hat{\mu}_*(t) < \mu_* - (|\Delta_*^S| + \beta_*(t))\big) \\ &\leqslant \sum_{t=1}^{T} \sum_{n=1}^{t} \mathbb{P}\left(\hat{\mu}_{*,n} < \mu_* - \left(|\Delta_*^S| + \sqrt{\tfrac{2\log(f(t))}{n}}\right)\right) \\ &\leqslant \sum_{t=1}^{T} \sum_{n=1}^{t} \frac{1}{f(t)} \exp\big(-\tfrac{n|\Delta_*^S|^2}{2}\big) \leqslant \frac{2}{|\Delta_*^S|^2} \sum_{t=1}^{T} \frac{1}{f(t)} \leqslant \frac{5}{|\Delta_*^S|^2}\,. \end{aligned} \tag{9}$$

The last inequality is obtained by observing that $\sum_{t=1}^{T} \frac{1}{f(t)} \leqslant 1 + \sum_{t=2}^{T} \frac{1}{t \log^2(t)}$ and then bounding the sum with an integral.

Finally, by putting everything together, we obtain from equations (6), (7), (8), and (9) the claimed result

$$R_T^S = \sum_{i:\Delta_i^S>0} \Delta_i^S \, \mathbb{E}(n_i(T)) \leqslant \sum_{i:\Delta_i^S>0} \left(\Delta_i^S + \frac{2}{\Delta_i^S} + \frac{7\Delta_i^S}{|\Delta_*^S|^2} \right). \qquad \square$$

Remark. As can be easiliy seen from the proof, instead of choosing the arm whose share of confidence interval above S is maximal (line 4 of Algorithm 3) it is sufficient to choose any arm with empirical mean above the satisfaction level. Obvious alternatives are choosing the arm with maximal empirical mean reward or using UCB1 to choose among the arms with empirical mean reward above S. We will however see that Algorithm 3 empirically outperforms these simpler alternatives.

Theorem 4. *If $\mu_* \leqslant S$ then Algorithm 3 satisfies for all $T \geqslant 1$*

$$R_T \leqslant \sum_{i:\Delta_i>0} \inf_{\varepsilon \in (0,\Delta_i)} \Delta_i \left(1 + \frac{5}{\varepsilon^2} + \frac{2(\log f(T) + \sqrt{\pi \log f(T)} + 1)}{(\Delta_i - \varepsilon)^2} \right). \tag{10}$$

Furthermore,

$$\limsup_{T\to\infty} \frac{R_T}{\log(T)} \leqslant \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}. \tag{11}$$

Thus, for a constant $C > 0$ it holds that

$$R_T \leqslant C \sum_{i:\Delta_i>0} \left(\Delta_i + \frac{\log(T)}{\Delta_i} \right).$$

Proof. The proof can be reduced to the derivation of the regret bounds for UCB1 as given in Theorem 8.1 of Lattimore and Szepesvári (2020). We start with the standard regret decomposition

$$R_T = \sum_{i:\Delta_i>0} \mathbb{E}\left(n_i(T)\right) \Delta_i.$$

In the following, we bound for each suboptimal arm i the number of times $n_i(T)$ it is played. Note that arm i is chosen after step T only if either

$$\hat{\mu}_i(t) + \beta_i(t) \geqslant \hat{\mu}_*(t) + \beta_*(t) \quad \text{or} \quad \hat{\mu}_i(t) + \beta_i(t) \geqslant S.$$

(Note that the case $\hat{\mu}_i(t) \geqslant S$ is subsumed by the second event.) Accordingly, we can decompose the event $A_t = i$ using some arbitrary but fixed $\varepsilon \in (0, \Delta_i)$ as

$$\begin{aligned}
\{A_t = i\} \subseteq & \left\{A_t = i \text{ and } \hat{\mu}_*(t) + \beta_*(t) \leqslant \mu_* - \varepsilon\right\} \cup \left\{A_t = i \text{ and } \hat{\mu}_*(t) + \beta_*(t) \geqslant \mu_* - \varepsilon\right\} \\
\subseteq & \left\{\hat{\mu}_*(t) + \beta_*(t) \leqslant \mu_* - \varepsilon\right\} \\
& \cup \left\{A_t = i \text{ and } \hat{\mu}_i(t) + \beta_i(t) \geqslant \hat{\mu}_*(t) + \beta_*(t) \geqslant \mu_* - \varepsilon\right\} \\
& \cup \left\{A_t = i \text{ and } \hat{\mu}_*(t) + \beta_*(t) \geqslant \mu_* - \varepsilon \text{ and } \hat{\mu}_i(t) + \beta_i(t) \geqslant S\right\} \\
\subseteq & \left\{\hat{\mu}_*(t) + \beta_*(t) \leqslant \mu_* - \varepsilon\right\} \\
& \cup \left\{A_t = i \text{ and } \hat{\mu}_i(t) + \beta_i(t) \geqslant \mu_* - \varepsilon\right\},
\end{aligned}$$

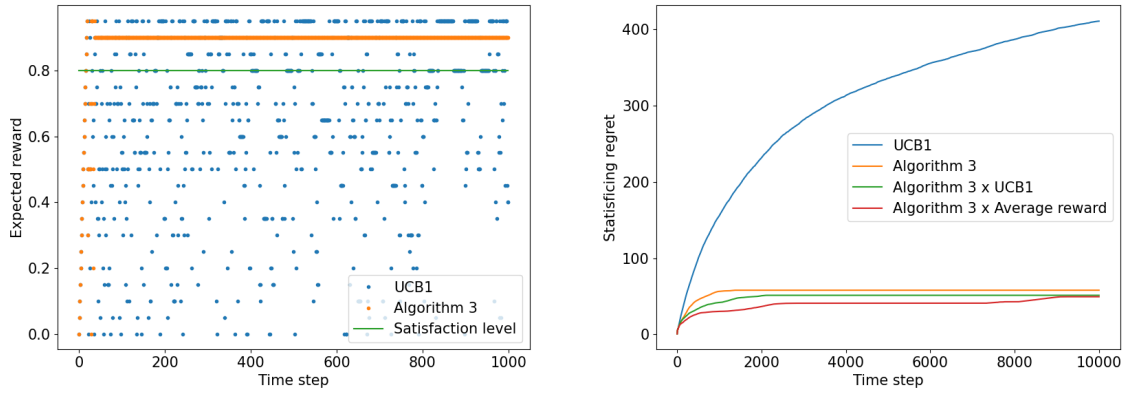
where the last inclusion is due to the assumption that $\mu_* \leqslant S$. It follows that

$$n_i(T) \leqslant \sum_{t=1}^{T} \mathbb{I}\big\{\hat{\mu}_*(t) + \beta_*(t) \leqslant \mu_1 - \varepsilon\big\} + \sum_{t=1}^{T} \mathbb{I}\big\{A_t = i \text{ and } \hat{\mu}_i(t) + \beta_i(t) \geqslant \mu_* - \varepsilon\big\}.$$

The obtained decomposition is the same as the one in the proof of Theorem 8.1 from (Lattimore and Szepesvári, 2020) and the very same arguments can be used to finish the proof of (10). The second part of the theorem, that is eq. (11), follows by choosing $\varepsilon = \log^{-1/4}(T)$ and taking the limit as T tends to infinity. □

5. Experiments

We compared our Algorithm 3 to standard bandit algorithms in order to show that the latter keep accumulating S-regret, while Algorithm 3 sticks to a satisficing arm after finite time, thus confirming the results of Theorem 3. We started with comparing Algorithm 3 to UCB1 (Auer et al., 2002) in a setting with 20 arms and normally distributed rewards with standard deviation 1 and the mean reward of arm i set to $\frac{i-1}{20}$. The satisfaction level was set to 0.8.



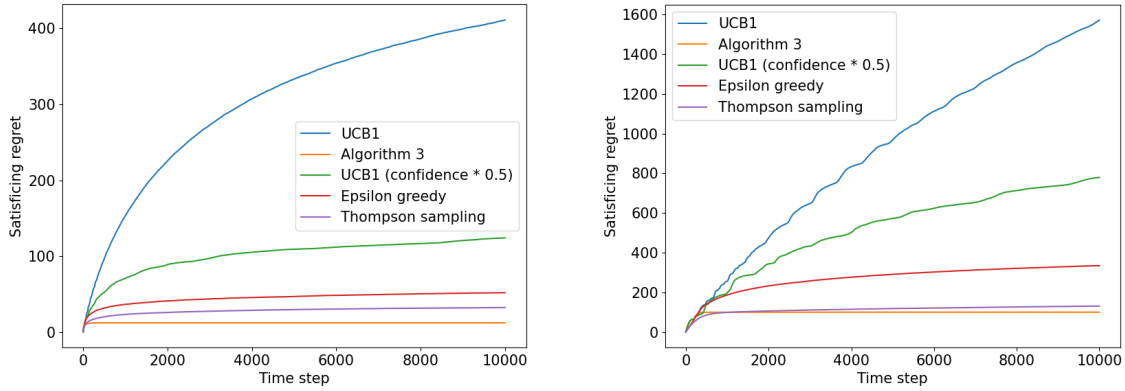
(a) Arm pulls for each algorithm in exemplary run.

(b) S-regret averaged over 50 runs.

Figure 1: Experiments with Gaussian bandits comparing Algorithm 3 to UCB1.

Figure 1a depicts a showcase run illustrating that Algorithm 3 soon focuses on a satisficing arm, while UCB1 keeps exploring. Figure 1b gives the S-regret averaged over 50 runs. Although the latter in general is smaller than classic regret, UCB1 suffers growing S-regret due to ongoing exploration of arms below the satisfaction level. Figure 1b also compares Algorithm 3 to variants that use a different criterion for choosing among empirically satisficing arms. That is, instead of using the fraction index in line 4 of Algorithm 3 we consider using UCB1 for choosing an arm or pick the arm having the highest empirical mean, respectively. We see in Figure 1b that the original version of Algorithm 3 works best.

Next, we considered Bernoulli distributed rewards and added Thompson sampling (Thompson, 1933) and ε-greedy (Auer et al., 2002) to the comparison. For ε-greedy we chose $\varepsilon_t := \frac{K}{10t}$ at each step t, while we used a version of Thompson sampling adapted to Bernoulli rewards, using Beta distributions for the estimate. Further, for UCB1 we halved the bonus term for focusing more on exploitation. Figure 2 shows that Algorithm 3 is still the only one giving constant S-regret. Not surprisingly, if the number of arms is raised from 20 to 200, S-regret increases. However, maybe with the exception of Thompson sampling the classic bandit algorithms seem to suffer more from the increase of the number of arms.



(a) S-regret for 20 arms averaged over 50 runs.

(b) S-regret for 200 arms averaged over 50 runs.

Figure 2: Experiments with Bernoulli bandits comparing Algorithm 3 to classic bandit algorithms.

Finally we also had a look at the not realizable case when the satisfaction level is chosen above μ_*. For this we used the same setup as before but chose $S = 1 > \mu_* = \frac{19}{20}$. Here the regret of the variants of Algorithm 3 is practically the same as for UCB1. Algorithm 3 itself performs a bit worse than UCB1 for some time, until the regret curve finally joins that of UCB1. Plots can be found in Figure 3 of Appendix B.

6. Conclusion

Our results for the multi-armed bandit case are just a first step in an ongoing project on satisficing in reinforcement learning. While some ideas may be used also in the general standard Markov decision process setting, it seems already not quite simple to obtain reasonable constant regret bounds in the realizable case. While it might be possible to consider each policy as an arm in a MAB setting, the resulting bounds would be linear in the number of policies and hence exponential in the number of states.

Also for the MAB setting itself, further improvements are possible. While we were happy to compete mainly with UCB1 in the non-realizable case, we are sure that suitable modifications of Algorithm 3 would give improved experimental performance while keeping logarithmic regret bounds.

A lesson to take from the MAB setting is that the savings from considering a satisficing instead of an optimizing objective –at least with respect to regret– is not that there are arms that need no exploration at all. Rather in the worst case (as always considered by notions of regret) one still has to explore all arms, however the amount of necessary exploration is now constant and independent of the horizon.

Acknowledgments

The authors would like to thank anonymous NeurIPS reviewers for valuable comments on an earlier version of this paper. In particular, we are grateful for one reviewer pointing out an error in the proof of Theorem 3 for a slightly different variant of Algorithm 3. This work was supported by the Austrian Science Fund (FWF): TAI 590-N.

References

Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—part I: I.i.d. rewards. *IEEE Trans. Autom. Control*, 32:968–976, 12 1987.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002.

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *COLT 2013 – The 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 122–134, 2013.

Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. PAC identification of a bandit arm relative to a reward quantile. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI 2017*, pages 1777–1783, 2017.

Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. PAC identification of many good arms in stochastic multi-armed bandits. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 991–1000, 2019.

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

Hideaki Kano, Junya Honda, Kentaro Sakamaki, Kentaro Matsuura, Atsuyoshi Nakamura, and Masashi Sugiyama. Good arm identification via bandit feedback. *Mach. Learn.*, 108(5):721–745, 2019.

Yu Kohno and Tatsuji Takahashi. A cognitive satisficing strategy for bandit problems. *Int. J. Parallel Emergent Distrib. Syst.*, 32(2):232–242, 2017.

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

Blake Mason, Lalit Jain, Ardhendu Tripathy, and Robert Nowak. Finding all ε-good arms in stochastic bandits. In *Advances in Neural Information Processing Systems 33, NeurIPS 2020*, 2020.

Nadav Merlis and Shie Mannor. Lenient regret for multi-armed bandits. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 8950–8957, 2021.

Paul Reverdy, Vaibhav Srivastava, and Naomi Ehrich Leonard. Modeling human decision making in generalized Gaussian multiarmed bandits. *Proc. IEEE*, 102(4):544–571, 2014.

Paul Reverdy, Vaibhav Srivastava, and Naomi Ehrich Leonard. Satisficing in multi-armed bandit problems. *IEEE Trans. Autom. Control.*, 62(8):3788–3803, 2017.

Paul Reverdy, Vaibhav Srivastava, and Naomi Ehrich Leonard. Corrections to "Satisficing in multiarmed bandit problems". *IEEE Trans. Autom. Control*, 66(1):476–478, 2021.

Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *J. Artif. Intell. Res.*, 48:67–113, 2013.

Daniel Russo and Benjamin Van Roy. Satisficing in time-sensitive bandit learning. *CoRR*, abs/1803.02855, 2018.

Akihiro Tamatsukuri and Tatsuji Takahashi. Guaranteed satisficing and finite regret: Analysis of a cognitive satisficing value function. *Biosystems*, 180:46–53, 2019.

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

Appendix A. Exploration Based on a Potential Function

Following Bubeck et al. (2013), Algorithm 2 presents a more general approach for exploration, using a potential function $\psi : [0, \infty) \to \mathbb{R}^+$ that is assumed to be differentiable and increasing.

Algorithm 2

Require: K, S

Play each arm once, i.e., for time steps $t = 1, \dots, K$ play arm $A_t = t$.
for time steps $t = K + 1, \dots$ **do**
  if $\exists i \, \hat{\mu}_i(t) \geqslant S$ **then**
    Play $A_t \leftarrow \operatorname{argmax}_{i \in [\![1,K]\!]} \hat{\mu}_i(t)$.
  else
    Choose randomly an arm according to the probability distribution defined by

$$p_{i,t} = \frac{1}{\alpha \times \psi(|S - \hat{\mu}_i(t)|)}, \text{ where } \alpha = \sum_{j=1}^{K} \frac{1}{\psi(|S - \hat{\mu}_j(t)|)}.$$

  end if
end for

Theorem 5. *Let $\psi : [0, \infty) \to \mathbb{R}^+$ be a differentiable and increasing function. If $\mu_* > S$ then Algorithm 2 satisfies for all $T \geqslant 1$,*

$$R_T^S \leqslant \sum_{i:\Delta_i^S > 0} \left(\Delta_i^S + \frac{8}{\Delta_i^S} + \frac{\Delta_i^S}{\psi(\frac{\Delta_i^S}{2})} \left(\frac{2\psi(0)}{(\Delta_*^S)^2} + \int_0^{+\infty} \frac{\psi'(x)}{e^{\frac{(|\Delta_*^S| + x)^2}{2}} - 1} \, dx \right) \right).$$

Proof. As in the proof of Theorem 1 we aim at a bound on $\mathbb{E}(n_i(T)) = \sum_{t=1}^{T} \mathbb{P}(A_t = i)$ for each non-satisficing arm i. First, we decompose the event $\{A_t = i\}$ as

$$\{A_t = i\} \subset \{t = i\} \cup \left\{A_t = i, \hat{\mu}_i(t) > S - \tfrac{\Delta_i^S}{2}, t > K\right\} \cup \left\{A_t = i, \hat{\mu}_i(t) \leqslant S - \tfrac{\Delta_i}{2}, t > K\right\}. \quad (12)$$

For the first two events we have

$$\sum_{t=1}^{T} \mathbb{P}(t = i) \leqslant 1 \quad \text{and} \quad \sum_{t=1}^{T} \mathbb{P}\left\{A_t = i, \hat{\mu}_i(t) > S - \tfrac{\Delta_i^S}{2}, t > K\right\} \leqslant \frac{8}{(\Delta_i^S)^2}.$$

For the probability of the third event in (12) we have

$$\begin{aligned}
\mathbb{P}\big(A_t = i, \hat{\mu}_i(t) \leqslant S - \tfrac{\Delta_i^S}{2}, t > K\big) &\leqslant \mathbb{P}\big(A_t = i, \hat{\mu}_i(t) \leqslant S - \tfrac{\Delta_i^S}{2}, Z_t\big) \\
&= \mathbb{E}\big(p_{i,t} \mathbb{1}\big\{\hat{\mu}_i(t) \leqslant S - \tfrac{\Delta_i^S}{2}, Z_t\big\}\big) \\
&= \mathbb{E}\Big(\frac{p_{i,t}}{p_{*,t}} p_{*,t} \mathbb{1}\big\{\hat{\mu}_i(t) \leqslant S - \tfrac{\Delta_i^S}{2}, Z_t\big\}\Big) \\
&\leqslant \mathbb{E}\Big(\frac{\psi(|S - \hat{\mu}_*(t)|)}{\psi\big(\frac{\Delta_i^S}{2}\big)} p_{*,t} \mathbb{1}\big\{\hat{\mu}_i(t) \leqslant S - \tfrac{\Delta_i^S}{2}, Z_t\big\}\Big) \\
&\leqslant \frac{1}{\psi\big(\frac{\Delta_i^S}{2}\big)} \mathbb{E}\big(\psi(|S - \hat{\mu}_*(t)|)\, p_{*,t}\, \mathbb{1}\{Z_t\}\big) \\
&\leqslant \frac{1}{\psi\big(\frac{\Delta_i^S}{2}\big)} \mathbb{E}\big(\psi(|S - \hat{\mu}_*(t)|)\, \mathbb{1}\{A_t = *, Z_t\}\big).
\end{aligned}$$

Summing up the expectation value over all t yields

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\big(\psi(|S - \hat{\mu}_*(t)|)\, \mathbb{1}\{A_t = *, Z_t\}\big) &\leqslant \sum_{t=1}^{T} \mathbb{E}\big(\psi(|S - \hat{\mu}_{*,t}|)\, \mathbb{1}\{\hat{\mu}_{*,t} \leqslant S\}\big) \\
&= \sum_{n=1}^{T} \int_0^{\infty} \mathbb{P}\big(\psi(|S - \hat{\mu}_{*,n}|)\, \mathbb{1}\{\hat{\mu}_{*,n} \leqslant S\} \geqslant x\big)\, dx \\
&= \sum_{n=1}^{T} \int_0^{\psi(0)} \mathbb{P}\big(\psi(|S - \hat{\mu}_{*,n}|)\, \mathbb{1}\{\hat{\mu}_{*,n} \leqslant S\} \geqslant x\big)\, dx \\
&\quad + \sum_{n=1}^{T} \int_{\psi(0)}^{\infty} \mathbb{P}\big(\psi(|S - \hat{\mu}_{*,n}|)\, \mathbb{1}\{\hat{\mu}_{*,n} \leqslant S\} \geqslant x\big)\, dx \\
&= \sum_{n=1}^{T} \int_0^{\psi(0)} \mathbb{P}\big(\hat{\mu}_{*,n} \leqslant S\big)\, dx + \sum_{n=1}^{T} \int_{\psi(0)}^{\psi(\infty)} \mathbb{P}\big(\hat{\mu}_{*,n} \leqslant S - \psi^{-1}(x)\big)\, dx,
\end{aligned}$$

noting that, since ψ is increasing, for $x \leqslant \psi(0)$ the inequality $\psi(S - \hat{\mu}_{*,n})\mathbb{1}\{\hat{\mu}_{*,n} \leqslant S\} \geqslant x$ is equivalent to $\mathbb{1}\{\hat{\mu}_{*,n} \leqslant S\} = 1$, while for $x \geqslant \psi(0)$ it is equivalent to $\psi(S - \hat{\mu}_{*,n}) \geqslant x$. Further, note that if $x > \psi(\infty) := \lim_{y\to\infty} \psi(y)$ then the integrand is equal to 0.

We continue with the analysis of the same term and obtain

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\big(\psi(|S - \hat{\mu}_*(t)|)\, \mathbb{1}\{A_t = *, Z_t\}\big) &\leqslant \sum_{n=1}^{T} \psi(0)\, \mathbb{P}\big(\hat{\mu}_{*,n} \leqslant S\big) + \sum_{n=1}^{T} \int_0^{\infty} \mathbb{P}\big(\hat{\mu}_{*,n} \leqslant S - u\big)\, \psi'(u)\, du \\
&\leqslant \sum_{n=1}^{T} \psi(0) \exp\big(-\tfrac{n(\Delta_*^S)^2}{2}\big) + \sum_{n=1}^{T} \int_0^{\infty} \exp\big(-\tfrac{n(|\Delta_*^S| + u)^2}{2}\big)\, \psi'(u)\, du \\
&\leqslant \frac{2\psi(0)}{(\Delta_*^S)^2} + \int_0^{\infty} \sum_{n=1}^{T} \exp\big(-\tfrac{n(|\Delta_*^S| + u)^2}{2}\big)\, \psi'(u)\, du \\
&\leqslant \frac{2\psi(0)}{(\Delta_*^S)^2} + \int_0^{\infty} \frac{\psi'(u)}{e^{\frac{(|\Delta_*^S| + u)^2}{2}} - 1}\, du.
\end{aligned}$$

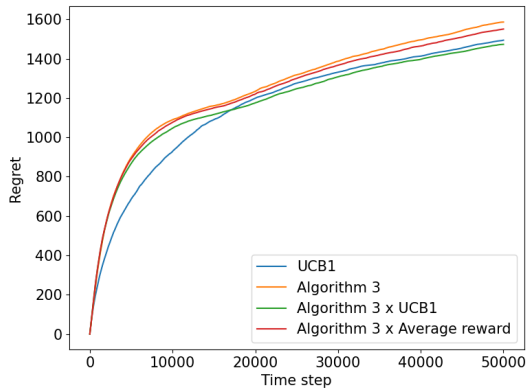
Finally, by putting everything together, we obtain

$$R_T^S = \sum_{i:\Delta_i^S>0} \mathbb{E}(n_i(T))\,\Delta_i^S$$
$$\leqslant \sum_{i:\Delta_i^S>0} \left(\Delta_i^S + \frac{8}{\Delta_i^S} + \frac{\Delta_i^S}{\psi\left(\frac{\Delta_i^S}{2}\right)}\left(\frac{2\psi(0)}{(\Delta_*^S)^2} + \int_0^\infty \frac{\psi'(x)}{e^{\frac{(|\Delta_*^S|+x)^2}{2}} - 1}\,dx\right)\right),$$

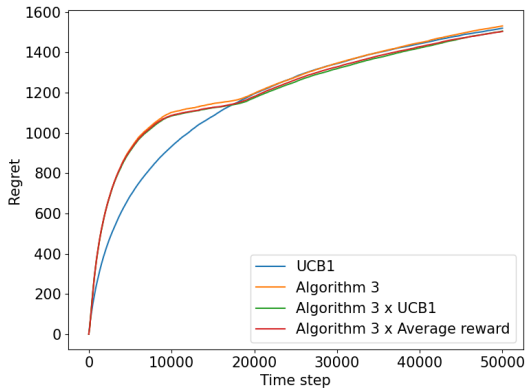
which completes the proof. □

As discussed by Bubeck et al. (2013) a simple choice for the potential function is $\psi(x) = x^2$, which gives a bound similar to that of Theorem 1. For refined choices for Ψ we refer to Section 3 of (Bubeck et al., 2013).

Appendix B. Plots for the Realizable Case



(a) Classic regret for Gaussian bandits averaged over 50 runs.



(b) Classic regret for Bernoulli bandits averaged over 50 runs.

Figure 3: Experiments for the not realizable case when $S > \mu_*$.