Analysis of Thompson Sampling for Stochastic Sleeping Bandits

Aritra Chatterjee, Ganesh Ghalme, Shweta Jain, Rohit Vaish and **Y. Narahari**
Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India
{aritra.chatterjee, ganesh.ghalme, jainshweta, rohit.vaish, hari}@csa.iisc.ernet.in

Abstract

We study a variant of the stochastic multi-armed bandit problem where the set of available arms varies arbitrarily with time (also known as the *sleeping bandit* problem). We focus on the Thompson Sampling algorithm and consider a regret notion defined with respect to the best *available* arm. Our main result is an $\mathcal{O}(\log T)$ regret bound for Thompson Sampling, which generalizes a similar bound known for this algorithm from the classical bandit setting. Our bound also matches (up to constants) the best-known lower bound for the sleeping bandit problem. We show via simulations that Thompson Sampling outperforms the UCB-style AUER algorithm for the sleeping bandit problem.

1 INTRODUCTION

In the classical *multi-armed bandit* (MAB) problem, an algorithm is required to choose one of the K available actions in each of the T rounds. Each choice of action generates a stochastic reward drawn from an unknown but fixed distribution. The goal of the algorithm is to maximize its expected sum of rewards over the T rounds, relative to the best-fixed action in hindsight (in other words, minimize its *regret*).

The MAB framework can model a variety of situations that involve sequential decision-making, such as online advertising, network routing, and cloud computing (refer to the survey by Bubeck and Cesa-Bianchi, 2012 for other applications). In many of these settings, however, not all of the actions are available at all stages. Indeed, some advertisers might choose to stay away from the auction due to budget constraints, some links might be unavailable due to congestion, and some servers might be unavailable due to maintenance - forcing the algorithm to work only with the set of *available* actions. Such problems fall under the *sleeping multi-armed bandit* (SMAB) model, which generalizes the standard MAB framework.

In the SMAB model, at each round, the algorithm is required to select an action from a set of *available* arms. Just like in the classical setting, the algorithm receives a stochastic reward based on the chosen action. However, unlike the classical setting, the performance of the algorithm can no longer be evaluated relative to a single fixed action, since each action might become unavailable at some stage of the algorithm. For this reason, the regret of an algorithm for SMAB (for a given set of availabilities of arms) is measured relative to the best *available* action in hindsight (refer to Section 3.1 for a formal definition). The overall regret of the algorithm is given by the worst-case regret over all settings of availabilities (i.e., under adversarial selection of the available arms).

The SMAB problem was first studied by Kleinberg et al. (2010) for both stochastic and adversarial rewards, and under adversarial availabilities. Their analysis focuses on the AUER algorithm, which is a variant of the well-known UCB1 algorithm for the classical MAB problem (Auer et al., 2002). Kleinberg et al. (2010) provide an $\mathcal{O}(\log T)$ bound on the regret of AUER and show that this is information-theoretically optimal (up to constants). However, their regret bound lacks a fine-grained dependence on the structure of availability of the arms.

Our interest in this paper is in studying the SMAB problem for a different algorithm called *Thompson Sampling* (Thompson, 1933). This algorithm has generated significant interest recently due to its impressive theoretical and empirical properties in the classical MAB setting (Chapelle and Li, 2011; Kaufmann et al., 2012; Agrawal and Goyal, 2012, 2013a). In this paper, we evaluate the performance of Thompson Sampling for the SMAB problem both from a theoretical and an empirical standpoint.

Our contributions

- On the theoretical side, we provide an $\mathcal{O}(\log T)$ regret bound for Thompson Sampling for the SMAB problem (Corollary 1). This bound generalizes the bound known for this algorithm from the classical setting (Kaufmann et al., 2012), and also matches (up to constants) the information-theoretic lower bound for the SMAB problem (Kleinberg et al., 2010). An interesting feature of our bound is that it provides a fine-grained dependence of the regret on the availabilities of the arms (Theorem 1).
- On the empirical side, we show through simulations on synthetic datasets that Thompson Sampling outperforms AUER algorithm for various settings of availabilities (Section 5). We also find that the performance of Thompson Sampling (across availabilities) is consistent with the prediction of our regret bound in Theorem 1.

2 RELATED WORK

The SMAB problem was first considered by Kleinberg et al. (2010). They formulate the notion of regret as the difference between the expected cumulative reward of algorithm and the expected reward from an optimal policy which picks arms according to the *best ordering* in hindsight. For this notion of regret, Kleinberg et al. (2010) proved an $\Omega(\log T)$ lower bound on regret for any algorithm for SMAB (Proposition 2), and provided a UCB-style algorithm (AUER) with a matching upper bound up to constants (Proposition 3). The notion of regret used in our work (defined in Section 3.1) is equivalent to that of Kleinberg et al. (2010) as in both cases the optimal policy picks the best among the available arms according to a fixed and unknown ordering. Follow-up work by Kanade et al. (2009) and Kanade and Steinke (2014) has studied settings where the rewards are adversarial while the availability of arms is either stochastic or adversarial.

The work of Kaufmann et al. (2012) on Thompson Sampling for the classical MAB setting is closest to ours. They provided an asymptotically-optimal regret bound - generalizing the previously-known logarithmic bound of Agrawal and Goyal (2012). Our regret bound can be seen as a generalization of this result to the sleeping MAB problem. However, unlike Kaufmann et al. (2012), our regret guarantee is not asymptotically-optimal.

Thompson Sampling can be adapted to several other variations of the classical MAB problem such as the budgeted MAB problem (Xia et al., 2015), multi-pull MAB problem (Komiyama et al., 2015) and contextual MAB problem (Agrawal and Goyal, 2013b). Several papers (Gentile et al., 2014; Li et al., 2016) have studied UCB-based algorithms for the contextual MAB setting, but Li and Chapelle (2012) showed through detailed simulations that Thompson Sampling performs competitively even in such settings.

3 PRELIMINARIES

3.1 THE MODEL

An instance of the *sleeping multi-armed bandit* (SMAB) problem is a tuple $\langle [K], \{\mu_i\}_{i\in[K]}, \{A_t\}_{t\in[T]}\rangle$, where $[K]$ denotes the set of arms $\{1, 2, \ldots, K\}$, $\mu_i \in (0, 1)$ denotes the Bernoulli parameter for arm i, $A_t \subseteq [K]$ denotes the set of available arms at time instant t, and the parameter $T \in \mathbb{N}$ denotes the *time horizon*. For simplicity, we will assume that $\mu_1 > \mu_2 > \cdots > \mu_K$. When $A_t = [K]$ at each time t, we recover the classical MAB model.

Table 1: Notation

Notation	Definition
$[K]$	Set of arms.
T	Time horizon (or the number of rounds).
A_t	Set of arms available at time t.
$\mathcal{A}$	Availability sequence $\{A_1, \ldots, A_T\}$.
μ_i	Bernoulli parameter (or mean) for arm i.
i_t	Arm pulled at time t.
r_t	Reward obtained at time t.
i^*_t	Best available (or optimal) arm at time t.
$\theta_{i,t}$	Thompson sample of arm i at time t.
$s_{i,t}$	Number of successful pulls of arm i until time t.
$n_{i,t}$	Number of pulls of arm i until time t.
$\Delta_{i,j}$	$\lvert\mu_i - \mu_j\rvert$.
$\delta_{i,j}$	$\Delta_{i,j}/2$.
$a_{i,t}$	Number of times arm i is available until time t.
$N_{i,j}(t)$	Number of times arm j is pulled until time t when some arm in $[i]$ is available.
$L_{i,j}$	$32/\Delta^2_{i,j}$.
$w_{i,t}$	Number of times arm i is optimal until time t.
$e_{i,t}$	$\sqrt{4\ln(w_{i,t})/n_{i,t}}$.
$Q_{i,j}$	$4\ln(w_{i,t})/\delta^2_{i,j}$.

An *algorithm* for SMAB is specified as follows: At each time instant t, the algorithm receives as input the set of available arms A_t. The algorithm then outputs the index of an arm $i_t \in A_t$, and observes a reward r_t drawn i.i.d. from a Bernoulli distribution with parameter μ_{i_t}. The arm i_t is said to be *pulled* by the algorithm at time t.

The performance of an algorithm for SMAB problem is measured in terms of its *regret*. In order to formalize this, let $\mathcal{A} := \{A_1, \ldots, A_T\}$ denote the sequence of the availability sets (we call $\mathcal{A}$ the *availability sequence*). For any time instant t, let $i_t \in A_t$ denote the arm pulled by the algorithm, and let $i_t^* \in \arg\max_{i \in A_t} \mu_i$ denote the best available (or *optimal*) arm. Then, the expected regret of the algorithm for a given availability sequence $\mathcal{A}$ is,

$$R_{\mathcal{A}}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \mu_{i_t^*} - \mu_{i_t}\right],$$

where the expectation is taken over the random choices $\{i_t\}_{t \in [T]}$ made by the algorithm. The overall expected regret $R(T)$ of the algorithm is defined as its expected regret for the worst-case availability sequence, i.e.,

$$R(T) = \max_{\mathcal{A}} R_{\mathcal{A}}(T).$$

From here onwards, we will call $R(T)$ the *regret* of the algorithm, and $R_{\mathcal{A}}(T)$ the *availability-specific regret*. We remark that the time horizon T is used only in the analysis of regret, and is not used by the algorithm in deciding which arm to pull. Table 1 provides a list of the key notation used in the paper.

3.2 THOMPSON SAMPLING ALGORITHM FOR SMAB

Algorithm 1: Thompson Sampling (TS-SMAB)

```
Input  : Set of arms $[K]$.
Output: Set of pulls $\{i_t\}_{t=1}^{T}$.
Initialize $s_{i,1} = 0$ and $n_{i,1} = 0$ for all $i \in [K]$.
for $t \leftarrow 1$ to $T$ do
    Observe the set of available arms $A_t \subseteq [K]$.
    Sample $\theta_{i,t} \sim$ Beta$(s_{i,t} + 1, n_{i,t} - s_{i,t} + 1)$ for
      each arm $i \in A_t$.
    Pull the arm $i_t \in \arg\max_{i \in A_t} \theta_{i,t}$.
      (ties are broken lexicographically)
    Observe reward $r_t \sim$ Bernoulli$(\mu_{i_t})$.
    Update posterior
        $s_{i_t,t+1} \leftarrow s_{i_t,t} + r_t$
        $n_{i_t,t+1} \leftarrow n_{i_t,t} + 1$
    for $j \neq i_t$ do
        $s_{j,t+1} \leftarrow s_{j,t}$
        $n_{j,t+1} \leftarrow n_{j,t}$
    end
end
```

In this work, we focus on the *Thompson Sampling* algorithm for the SMAB problem (denoted by TS-SMAB). Algorithm 1 presents a pseudo-code of TS-SMAB. At each time instant t, the algorithm observes the set of available arms A_t, and draws samples from Beta distributions for each available arm. The arm with the largest sample is pulled by the algorithm, and the Beta posteriors are updated based on the observed reward.

We remark that that the rewards are considered to be Bernoulli distributed throughout the paper. However, the algorithm and regret analysis can be adapted to more general reward distributions with support $[0, 1]$ by the simple extension provided by Agrawal and Goyal (2012) (Algorithm 2).

4 REGRET ANALYSIS

Our main result is a logarithmic bound on the availability-specific regret of TS-SMAB (Theorem 1). This bound is stated in terms of $a_{j,T}$, which is the number of times arm j is available until time T, i.e., $a_{j,T} = |\{t \in [T] : j \in A_t\}|$. This directly leads to an $\mathcal{O}(\log T)$ regret bound for TS-SMAB in Corollary 1.

Theorem 1. *The availability-specific regret of TS-SMAB is given by*

$$R_{\mathcal{A}}(T) \leq \sum_{i<j} \frac{32 \ln(a_{j,T})}{\Delta_{i,j}^2} \cdot \Delta_{i,i+1} + \mathcal{O}(1),$$

where $\Delta_{i,j} = |\mu_i - \mu_j|$ denotes the absolute difference of the mean rewards of the arms i and j.

In order to facilitate comparison with the existing results, we will make two simplifications to the bound in Theorem 1: First, we will upper bound the term $a_{j,T}$ by T (thereby removing the dependence on $\mathcal{A}$). Second, we will use Proposition 1 below to rewrite the dependence on the Δ parameters.

Proposition 1 (Kleinberg et al., 2010)**.**

$$\sum_{j=2}^{K} \sum_{i=1}^{j-1} \Delta_{i,j}^{-2} \Delta_{i,i+1} \leq 2 \sum_{i=1}^{K-1} \Delta_{i,i+1}^{-1}.$$

Corollary 1. *The regret of TS-SMAB is given by*

$$R(T) \leq 64 \ln(T) \cdot \sum_{i=1}^{K-1} \frac{1}{\Delta_{i,i+1}} + \mathcal{O}(1).$$

We can now compare our regret bound in Corollary 1 with the known results. First, we observe that our bound is *information-theoretically optimal*, i.e., it matches (up to problem-independent constants) the following lower bound for SMAB problem due to Kleinberg et al. (2010):

Proposition 2 (Kleinberg et al., 2010). *Let $\mu_i \in (a, b)$ for all $i \in [K]$ and some $0 < a < b < 1$, s.t. $\mu_1 > \mu_2 > \ldots > \mu_K$. Then, the regret of any algorithm for SMAB is at least*

$$\Omega\left(\ln(T) \cdot \sum_{i=1}^{K-1} \frac{1}{\Delta_{i,i+1}}\right).$$

Next, we remark that our bound in Corollary 1 for TS-SMAB is identical to that of the AUER algorithm for SMAB.

Proposition 3 (Kleinberg et al., 2010). *The regret of AUER algorithm for SMAB is given by*

$$R(T) \leq 64 \ln(T) \cdot \sum_{i=1}^{K-1} \frac{1}{\Delta_{i,i+1}} + \mathcal{O}(1).$$

Finally, our bound in Corollary 1 matches the bound for Thompson Sampling for *classical* MAB setting up to problem-dependent constants.

Proposition 4 (Kaufmann et al., 2012). *Let $A_t = [K]$ for all $t \in [T]$. Then, the availability-specific regret of TS-SMAB is given by*

$$R_{\mathcal{A}}(T) \leq \sum_{i=2}^{K} \frac{\ln(T) + \ln\ln(T)}{\mathrm{KL}(\mu_i, \mu_1)} + \mathcal{O}(1).$$

4.1 PROOF OF MAIN RESULT (THEOREM 1)

Our proof for the regret bound in Theorem 1 relies on a key lemma (Lemma 1), which bounds the expected number of *suboptimal* pulls of any arm. A suboptimal pull is said to occur at time t if the algorithm pulls an arm $j \in A_t$ when a better arm i is available, i.e., $i_t = j$ and $\mu_i > \mu_j$. We use $N_{i,j}(t)$ to denote the number of times a fixed arm j is pulled until time t whenever some arm in the set $\{1, \ldots, i\}$ is available, i.e., $N_{i,j}(t) = |\{t' \in [T] : t' \leq t, i_{t'} = j, A_{t'} \cap [i] \neq \emptyset\}|$.

Lemma 1 (Bound on suboptimal pulls). *For any pair of arms i and j such that $\mu_i > \mu_j$, we have*

$$\mathbb{E}[N_{i,j}(T)] \leq L_{i,j} \ln(a_{j,T}) + \mathcal{O}(1),$$

where $L_{i,j} = 32/\Delta_{i,j}^2$.

We will now present the proof of Theorem 1, followed by the proof of Lemma 1.

Proof of Theorem 1. We reduce the problem of upper-bounding the availability-specific regret to upper-bounding the expected number of suboptimal pulls.

$$\begin{aligned}
R_{\mathcal{A}}(T) &= \mathbb{E}\left[\sum_{i<j} (N_{i,j} - N_{i-1,j}) \Delta_{i,j}\right] \\
&= \mathbb{E}\left[\sum_{i<j} N_{i,j} (\Delta_{i,j} - \Delta_{i+1,j})\right] \\
&\qquad \text{(we follow the convention } N_{0,j} = 0\text{)} \\
&= \sum_{i<j} \Delta_{i,i+1} \cdot \mathbb{E}[N_{i,j}] \\
&\leq \sum_{i<j} \Delta_{i,i+1} \cdot [L_{i,j} \ln(a_{j,T}) + \mathcal{O}(1)] \\
&\qquad \text{(Using Lemma 1)} \\
&= \sum_{i<j} \frac{32 \ln(a_{j,T})}{\Delta_{i,j}^2} \cdot \Delta_{i,i+1} + \mathcal{O}(1). \qquad \square
\end{aligned}$$

4.2 PROOF OF KEY LEMMA (LEMMA 1)

Our proof of Lemma 1 makes use of three intermediate results—Lemmas 2 to 4. Below, we state each of these results, followed by the proof of Lemma 1.

Our first two results (Lemmas 2 and 3) formalize the idea that pulling an arm j sufficiently many times results in the concentration of its Thompson samples around the mean. The proofs of these results follow from standard concentration arguments, and are deferred to the appendix (Sections B.1 and B.2).

Lemma 2. *Let $i, j \in [K]$ be a pair of arms such that $i < j$, and let $L_{i,j} = 32/\Delta_{i,j}^2$. Then, at any time $t \leq T$,*

$$\mathbb{P}(\theta_{j,t} \geq \mu_j + \Delta_{i,j}/2, n_{j,t} \geq L_{i,j} \ln(a_{j,T})) \leq 2a_{j,T}^{-3}.$$

Lemma 3. *Let $e_{i,t} = \sqrt{\frac{4\ln(w_{i,t})}{n_{i,t}}}$ for any arm i. Then, at any time t, we have,*

$$\mathbb{P}(\theta_{i,t} \leq \mu_i - e_{i,t}) \leq w_{i,t}^{-2},$$

where $w_{i,t}$ denotes the number of times arm i is the best available arm until time t.

Our next result (Lemma 4) is at the heart of our analysis, and is also technically the most involved. It associates the number of "optimal appearances" of an arm (i.e., the time instants at which a given arm is the best available arm) with the number of times it is pulled by the algorithm. More formally, for a given arm $i \in [K]$, let $w_{i,t}$ denote the number of time instants (until time t) where i is the best available arm. Then, Lemma 4 states that the number of times arm i is pulled by TS-SMAB (denoted by $n_{i,t}$) cannot be much smaller than $w_{i,t}$.

Lemma 4. *For each arm $i \in [K]$, there exist constants $b \in (0, 1)$ and $C_{i,b} < \infty$ such that*

$$\sum_{t \geq 1} \mathbb{P}\left(n_{i,t} \leq w_{i,t}^b\right) \leq C_{i,b}.$$

The proof of Lemma 4 extensively uses the techniques of Kaufmann et al. (2012), and is covered in Section 4.3. We now present the proof of Lemma 1 using Lemmas 2 to 4.

Proof of Lemma 1. Let $\zeta_{i,j}$ denote the set of all time instants until T for which the set of available arms includes the arm j and some arm in the set $\{1, \ldots, i\}$, i.e., $\zeta_{i,j} = \{t \in [T] : A_t \cap [i] \neq \emptyset, j \in A_t\}$. Thus, $|\zeta_{i,j}| \geq N_{i,j}(T)$. We therefore have

$$
\begin{aligned}
&\mathbb{E}[N_{ij}(T)] \\
&\leq \sum_{t \in \zeta_{i,j}} \mathbb{P}(i_t = j, A_t \cap [i] \neq \emptyset) \\
&\leq \sum_{t \in \zeta_{i,j}} \mathbb{P}(i_t = j, A_t \cap [i] \neq \emptyset, n_{j,t} \geq L_{i,j} \ln(a_{j,T})) \\
&\quad + L_{i,j} \ln(a_{j,T}). \qquad (1)
\end{aligned}
$$

The first term on the right can be analyzed as follows:

$$
\begin{aligned}
&\sum_{t \in \zeta_{i,j}} \mathbb{P}(i_t = j, A_t \cap [i] \neq \emptyset, n_{j,t} \geq L_{i,j} \ln(a_{j,T})) \\
&\leq \sum_{t \in \zeta_{i,j}} \mathbb{P}\left(\underbrace{\theta_{j,t} \geq \theta_{h_t,t}}_{E_1}, \underbrace{n_{j,t} \geq L_{i,j} \ln(a_{j,T})}_{E_2}\right),
\end{aligned}
$$

where $h_t := \arg\max_{k \in [i] \cap A_t} \mu_k$ denotes the best available arm in the set $\{1, \ldots, i\}$, and the event E_1 follows from the semantics of Thompson Sampling. We now use

$$
\begin{aligned}
\mathbb{P}(E_1 E_2) &= \mathbb{P}(E_1 E_2 E_3) + \mathbb{P}(E_1 E_2 E_3^{\mathsf{c}}) \\
&\leq \mathbb{P}(E_1 E_3) + \mathbb{P}(E_2 E_3^{\mathsf{c}}).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\sum_{t \in \zeta_{i,j}} \mathbb{P}(i_t = j, A_t \cap [i] \neq \emptyset, n_{j,t} \geq L_{i,j} \ln(a_{j,T})) \\
&\leq \sum_{t \in \zeta_{i,j}} \mathbb{P}\left(\underbrace{\theta_{j,t} \geq \theta_{h_t,t}}_{E_1}, \underbrace{\mu_j + \delta_{i,j} > \theta_{j,t}}_{E_3}\right) \\
&\quad + \sum_{t \in \zeta_{i,j}} \mathbb{P}\left(\underbrace{\theta_{j,t} \geq \mu_j + \delta_{i,j}}_{E_3^{\mathsf{c}}}, \underbrace{n_{j,t} \geq L_{i,j} \ln(a_{j,T})}_{E_2}\right), \\
&\qquad\qquad \text{(where } \delta_{i,j} = \Delta_{i,j}/2\text{)} \\
&\leq \sum_{t \in \zeta_{i,j}} \mathbb{P}(\mu_j + \delta_{i,j} \geq \theta_{h_t,t}) + \sum_{t \in \zeta_{i,j}} 2a_{j,T}^{-3} \\
&\qquad\qquad \text{(Using Lemma 2)} \\
&\leq \sum_{t \in \zeta_{i,j}} \mathbb{P}(\mu_j + \delta_{i,j} \geq \theta_{h_t,t}, \theta_{h_t,t} > \mu_{h_t} - e_{h_t,t}) \\
&\quad + \sum_{t \in \zeta_{i,j}} \mathbb{P}(\theta_{h_t,t} \leq \mu_{h_t} - e_{h_t,t}) + \sum_{i=1}^{a_{j,T}} 2a_{j,T}^{-3} \\
&\leq \sum_{t \in \zeta_{i,j}} \mathbb{P}(\mu_j + \delta_{i,j} \geq \mu_{h_t} - e_{h_t,t}) + \sum_{t \in \zeta_{i,j}} w_{h_t,t}^{-2} + 2a_{j,T}^{-2} \\
&\qquad\qquad \text{(Using Lemma 3)} \\
&\leq \sum_{t \in \zeta_{i,j}} \mathbb{P}(\mu_j + \delta_{h_t,j} \geq \mu_{h_t} - e_{h_t,t}) + \sum_{t \in \zeta_{i,j}} w_{h_t,t}^{-2} + \mathcal{O}(1) \\
&\leq \sum_{t \in \zeta_{i,j}} \mathbb{P}(-\delta_{h_t,j} \geq -e_{h_t,j}) + \sum_{t \in \zeta_{i,j}} w_{h_t,t}^{-2} + \mathcal{O}(1) \\
&\leq \sum_{t \in \zeta_{i,j}} \mathbb{P}(n_{h_t,t} \leq Q_{h_t,j}) + \sum_{k \leq i} \sum_{w_{k,t} \geq 1} w_{k,t}^{-2} + \mathcal{O}(1) \\
&\qquad\qquad (Q_{h_t,j} := \tfrac{4 \ln(w_{h_t,t})}{\delta_{h_t,j}^2}.) \\
&\leq \sum_{t \in \zeta_{i,j}} \mathbb{P}\Big(n_{h_t,t} \leq Q_{h_t,j}, Q_{h_t,j} \leq w_{h_t,t}^b\Big) \\
&\quad + \sum_{t \in \zeta_{i,j}} \mathbb{P}\Big(n_{h_t,t} \leq Q_{h_t,j}, Q_{h_t,j} > w_{h_t,t}^b\Big) + \mathcal{O}(1) \\
&\leq \sum_{t \in \zeta_{i,j}} \mathbb{P}\Big(n_{h_t,t} \leq w_{h_t,t}^b\Big) + \sum_{t \in \zeta_{i,j}} \mathbb{P}\Big(Q_{h_t,j} > w_{h_t,t}^b\Big) \\
&\quad + \mathcal{O}(1) \\
&\leq \sum_{k \leq i} \sum_{w_{k,t} \geq 1} \mathbb{P}\Big(n_{k,t} \leq w_{k,t}^b\Big) + \sum_{k \leq i} \sum_{w_{k,t} \geq 1} \mathbb{P}\Big(Q_{k,j} > w_{k,t}^b\Big) \\
&\quad + \mathcal{O}(1) \\
&\leq \sum_{k \leq i} C_{k,b} + \mathcal{O}(1) + \mathcal{O}(1) \qquad \text{(Using Lemma 4)} \\
&= \mathcal{O}(1).
\end{aligned}
$$

The $\mathcal{O}(1)$ bound on the term $\mathbb{P}\Big(Q_{k,j} > w_{k,t}^b\Big)$ follows from the observation that $Q_{k,j}$ is logarithmic in $w_{k,t}$, while $w_{k,t}^b$ is a polynomial. Along with Equation (1), the above bound gives the desired result. □

4.3 PROOF OF LEMMA 4

Outline of the proof: The description of the proof is made convenient by defining three different timescales. We use the term *original timeline* to refer to the time instants $t = 1, 2, \ldots$ and so on. The top-row in Figure 1 shows the original timeline. For a fixed arm i, let *i-awake timeline* denote the set of time instants for which arm i is available, shown as the middle row in Figure 1. Let $\tau_j^{(i)}$ denote the time instant (according to the original timeline) at which arm i is pulled for the j^{th} time (where $\tau_0^{(i)} := 0$). Consider the time interval (on the i-awake timeline) between the j^{th} and $(j+1)^{\text{th}}$ pulls of arm i (i.e., between $\tau_j^{(i)}$ and $\tau_{j+1}^{(i)}$). For this interval, let $\xi_j^{(i)}$ denote the set of time instants when arm i is the best available arm, i.e., $\xi_j^{(i)} = \Big\{t \in [T] : \tau_j^{(i)} < t < \tau_{j+1}^{(i)}, i \in A_t, A_t \cap [i-1] = \emptyset\Big\}$.

We refer to the set $\xi_j^{(i)}$ as the *i-optimal timeline*. Hence, the three timelines defined above are zoomed in (or

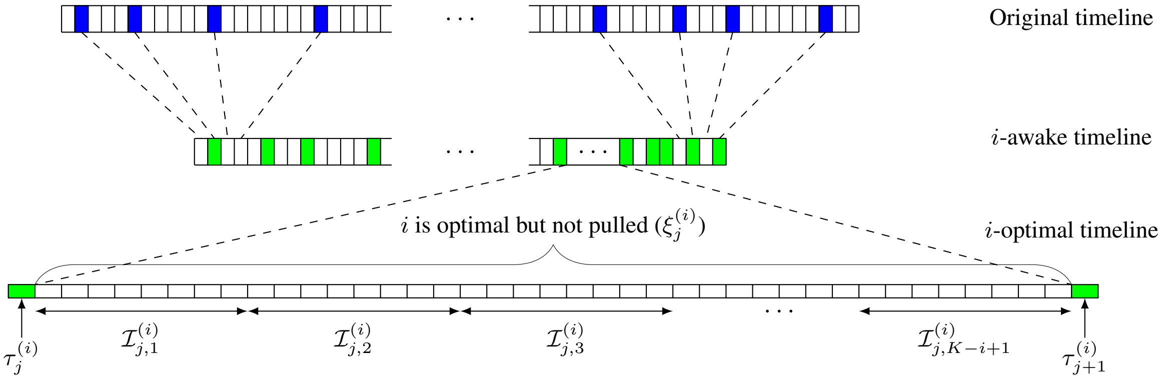
Original timeline
i-awake timeline
i is optimal but not pulled ($\xi_j^{(i)}$)
i-optimal timeline
$\tau_j^{(i)}$
$\mathcal{I}_{j,1}^{(i)}$
$\mathcal{I}_{j,2}^{(i)}$
$\mathcal{I}_{j,3}^{(i)}$
$\mathcal{I}_{j,K-i+1}^{(i)}$
$\tau_{j+1}^{(i)}$

Figure 1: The top row shows the original timeline, and the highlighted squares denote the time instants when arm i is available. The middle row shows the i-awake timeline (i.e., time instants when arm i is always available), and the highlighted squares represent the pulls of arm i. The bottom row zooms into the time instants between consecutive pulls of arm i, and shows the set of time instants where arm i is the best available arm but not pulled.

zoomed out) versions of each other.

Our goal in the proof will be to show that the i-optimal timeline cannot be too long. In other words, we will show that there cannot be a long run of time instants where arm i is the best available arm but is not pulled by the algorithm. This, in turn, will help us argue that the number of pulls of arm i (namely, $n_{i,t}$) cannot "lag behind" the number of optimal appearances of arm i (namely, $w_{i,t}$), giving us the desired result.

More specifically, we define $E_j^{(i)}$ as the event that the length of the i-optimal timeline ($|\xi_j^{(i)}|$) is at least $w_{i,t}^{1-b} - 1$. Under this event, starting from the instant $\tau_j^{(i)}$, we divide the i-optimal timeline into $(K-i+1)$ intervals, denoted by $\mathcal{I}_{j,l}^{(i)}$ for $l \in [K-i+1]$, each of length $\left\lceil \frac{w_{i,t}^{1-b}-1}{K-i+1} \right\rceil$ for some constant $b \in (0,1)$. (Hence, it is possible that the tail-end of the i-optimal timeline is not covered by these intervals). This division is shown in the bottom row in Figure 1.

Our proof is structured as follows:

- Step 1 of the proof reduces the problem to analyzing the event $E_j^{(i)}$.
- Steps 2 and 3 analyze the two mutually disjoint and exhaustive events, namely $E_j^{(i)} \cap F_{j,l}^{(i)}$ and $E_j^{(i)} \cap \overline{F_{j,l}^{(i)}}$, obtained via decomposition of the event $E_j^{(i)}$. The former describes the event where all suboptimal actions have been played sufficiently many times by the end of the interval $\mathcal{I}_{j,K-i}^{(i)}$, so that their samples are well-concentrated. Therefore, the probability of pulling a suboptimal arm during the interval $\mathcal{I}_{j,K-i+1}^{(i)}$ is very small. The latter describes the event that some suboptimal arm has not been played sufficiently often. For this case, we show that at the end of each interval $\mathcal{I}_{j,l}^{(i)}$, the algorithm must sample some (underexplored) suboptimal arm sufficiently many times. Since there are at least as many intervals as there are suboptimal arms, this gives a probability bound on the second event as well.

It is worth pointing out that the above proof outline closely follows the arguments of Kaufmann et al. (2012), who consider the analysis of Thompson Sampling in the classical MAB setting. The important difference lies in the conceptualization of the i-awake and the i-optimal timelines: In the model of Kaufmann et al. (2012), arm 1 is always available, hence analysis with respect to the original timeline suffices. In the SMAB problem, however, this is no more the case, and therefore we need to make the above arguments over different timescales.

Proof of Lemma 4. Let us fix an arm i. For any arm $a > i$, let $n_{a,t}^{(i)}$ denote the number of pulls of arm a over all time instants until time t whenever i is the best available arm. Notice that $\sum_{a>i} n_{a,t}^{(i)} = \sum_{j=0}^{n_{i,t}} \left|\xi_j^{(i)}\right|$.

Step 1: Reducing the problem to the analysis of $E_j^{(i)}$.

$$\begin{aligned}
&\mathbb{P}\left(n_{i,t} \le w_{i,t}^b\right) \\
&\le \mathbb{P}\left(n_{i,t}^{(i)} \le w_{i,t}^b\right) = \mathbb{P}\left(\sum_{a>i} n_{a,t}^{(i)} \ge w_{i,t} - w_{i,t}^b\right) \\
&\le \mathbb{P}\left(\exists j \in \left\{0, \ldots, \lfloor w_{i,t}^b \rfloor\right\} : \left|\xi_j^{(i)}\right| \ge w_{i,t}^{1-b} - 1\right) \\
&\le \sum_{j=0}^{\lfloor w_{i,t}^b \rfloor} \mathbb{P}\left(\underbrace{\left|\xi_j^{(i)}\right| \ge w_{i,t}^{1-b} - 1}_{E_j^{(i)}}\right).
\end{aligned} \tag{2}$$

Next, we formalize the idea of sampling a suboptimal arm "sufficiently many times". Given an arm a (where $a > i$), we say that a is *saturated* at time t if it has been pulled at least $L_{i,a}\ln(w_{i,t})$ times whenever arm i is the best available arm, i.e., $n_{a,t}^{(i)} \geq L_{i,a}\ln(w_{i,t})$. Otherwise, we say that arm a is *unsaturated*, and the act of pulling such an unsaturated arm is called an *interruption*.

Based on the above definitions, let us define $F_{j,l}^{(i)}$ to be the event that at least l suboptimal actions are saturated by the end of the interval $\mathcal{I}_{j,l}^{(i)}$. We also let $\gamma_{j,l}^{(i)}$ denote the number of interruptions during $\mathcal{I}_{j,l}^{(i)}$. A simple decomposition of $E_j^{(i)}$ gives us

$$\mathbb{P}\left(E_j^{(i)}\right) = \mathbb{P}\left(E_j^{(i)} \cap F_{j,K-i}^{(i)}\right) + \mathbb{P}\left(E_j^{(i)} \cap \overline{F_{j,K-i}^{(i)}}\right). \quad (3)$$

We analyze the two terms separately in Steps 2 and 3.

Step 2: In this step, we analyze the first term in Equation (3). Observe that the event $E_j^{(i)} \cap F_{j,K-i}^{(i)}$ implies that only the saturated arms are pulled during $\mathcal{I}_{j,K-i+1}^{(i)}$. Let $A := \{E_j^{(i)} \cap F_{j,K-i}^{(i)}\}$. Then,

$$\begin{aligned}
&\mathbb{P}\left(E_j^{(i)} \cap F_{j,K-i}^{(i)}\right) \\
&\leq \mathbb{P}\left(\left\{\exists t' \in \mathcal{I}_{j,K-i+1}^{(i)}, a > i : \theta_{a,t'} > \mu_a + \delta_{i,a}\right\} \cap A\right) \\
&\quad + \mathbb{P}\left(\left\{\forall t' \in \mathcal{I}_{j,K-i+1}^{(i)}, a > i : \theta_{a,t'} \leq \mu_a + \delta_{i,a}\right\} \cap A\right) \\
&\leq \mathbb{P}\Big(\exists t' \in \mathcal{I}_{j,K-i+1}^{(i)}, a > i : \theta_{a,t'} > \mu_a + \delta_{i,a}, \\
&\qquad\qquad n_{a,t}^{(i)} > L_{i,a}\ln(w_{i,t})\Big) \\
&\quad + \mathbb{P}\left(\left\{\forall t' \in \mathcal{I}_{j,K-i+1}^{(i)}, a > i : \theta_{a,t'} \leq y_i\right\} \cap A\right). \\
&\qquad\qquad (\text{where } y_i = \mu_{i+1} + \delta_{i,i+1})
\end{aligned}$$

We now use the following two results (Lemmas 5 and 6) whose proofs are deferred to the appendix (Sections B.3 and B.4).

Lemma 5. *For any fixed arm $a > i$,*

$$\mathbb{P}\left(\exists\, t' \in \mathcal{I}_{j,l}^{(i)} : \theta_{a,t'} \geq \mu_a + \delta_{i,a}, n_{a,t'}^{(i)} \geq L_{i,a}\ln(w_{i,t})\right) \leq 2w_{i,t}^{-2+b}.$$

Lemma 6. *There exists a constant $\lambda_0 > 1$ such that for all $\lambda \in (1, \lambda_0)$, for any interval $\mathcal{J}^{(i)} \subseteq \cup_{l\in[K-i+1]}\mathcal{I}_{j,l}^{(i)}$ and for every positive function f, we have*

$$\begin{aligned}
&\mathbb{P}\left(\{\forall s \in \mathcal{J}^{(i)} : \theta_{i,s} \leq y_i\} \cap \left\{\left|\mathcal{J}^{(i)}\right| \geq f(t)\right\}\right) \\
&\leq (\alpha_i)^{f(t)} + C_{\lambda,i}\frac{1}{f(t)^\lambda}e^{-jd_{\lambda,i}},
\end{aligned}$$

where $C_{\lambda,i} > 0$, $d_{\lambda,i} > 0$, $\alpha_i = \left(\frac{1}{2}\right)^{1-\mu_{i+1}-\delta_i}$ and $y_i = \mu_{i+1} + \delta_{i,i+1}$ for every $i \in \{1,\ldots,K-1\}$.

Using Lemmas 5 and 6, we get

$$\begin{aligned}
&\mathbb{P}\left(E_j^{(i)} \cap F_{j,K-i}^{(i)}\right) \\
&\leq \frac{2}{w_{i,t}^{2+b}} + (\alpha_i)^{\frac{w_{i,t}^{1-b}-1}{K-i+1}} + C_{\lambda,i}\left(\frac{w_{i,t}^{1-b}-1}{K-i+1}\right)^{-\lambda}e^{-jd_{\lambda,i}} \\
&= \frac{2}{w_{i,t}^{2+b}} + g(i,b,j,t).
\end{aligned}$$

Thus, $b < 1 - \frac{1}{\lambda} \Rightarrow \sum_{w_{i,t}\geq 1}\sum_{j\leq w_{i,t}^b} g(i,j,b,t) < +\infty$.

Step 3: This step provides the analysis for the second term in Equation (3). We claim that for any $2 \leq l \leq K$ and any time instant t greater than some constant $N_{i,b}$,

$$\mathbb{P}\left(E_j^{(i)} \cap \overline{F_{j,l-1}^{(i)}}\right) \leq (l-2)\left(\frac{2}{w_{i,t}^{2+b}} + f(i,b,j,t)\right).$$

We prove the above through induction over l. First, for the base case, observe that there exists a constant $N_{i,b}$ such that $w_{i,t} \geq N_{i,b} \Rightarrow \left\lceil\frac{w_{i,t}^{1-b}-1}{(K-i+1)^2}\right\rceil \geq L^{(i)}\ln(w_{i,t})$, (where $L^{(i)} = \max_{j>i} L_{i,j} = L_{i,i+1}$) which implies that at least one arm will be saturated by the end of interval $\mathcal{I}_{j,1}^{(i)}$. Hence for $w_{i,t} \geq N_{i,b}$, we have $\mathbb{P}\left(E_j^{(i)} \cap \overline{F_{j,1}^{(i)}}\right) = 0$. Next, we assume that the induction hypothesis holds for some $2 \leq l \leq K-i$.

Then,

$$\begin{aligned}
&\mathbb{P}\left(E_j^{(i)} \cap \overline{F_{j,l}^{(i)}}\right) \\
&\leq \mathbb{P}\left(E_j^{(i)} \cap \overline{F_{j,l-1}^{(i)}}\right) + \mathbb{P}\left(E_j^{(i)} \cap \overline{F_{j,l}^{(i)}} \cap F_{j,l-1}^{(i)}\right) \\
&\leq (l-2)\left(\frac{2}{w_{i,t}^{2+b}} + f(i,b,j,t)\right) \\
&\quad + \mathbb{P}\left(E_j^{(i)} \cap \overline{F_{j,l}^{(i)}} \cap F_{j,l-1}^{(i)}\right).
\end{aligned}$$

Hence, we need to show that

$$\mathbb{P}\left(E_j^{(i)} \cap \overline{F_{j,l}^{(i)}} \cap F_{j,l-1}^{(i)}\right) \leq \frac{2}{w_{i,t}^{2+b}} + f(i,b,j,t).$$

Observe that the event $E_j^{(i)} \cap \overline{F_{j,l}^{(i)}} \cap F_{j,l-1}^{(i)}$ implies that exactly $(l-1)$ arms are saturated at the beginning of interval $\mathcal{I}_{j,l}^{(i)}$, and no new arm is saturated during this interval. Hence, no unsaturated arm can be pulled more than $L^{(i)}\ln(w_{i,t})$ times, and thus, there cannot be more than $L^{(i)}(K-i)\ln(w_{i,t})$ interruptions during this interval. Let $\mathcal{S}_l^{(i)}$ denote the set of saturated arms at the end of $\mathcal{I}_{j,l}^{(i)}$. Then,

$$\begin{aligned}
&\mathbb{P}\Big(E_j^{(i)} \cap \overline{F_{j,l}^{(i)}} \cap F_{j,l-1}^{(i)}\Big)\\
&\le \mathbb{P}\Big(E_j^{(i)} \cap F_{j,l-1}^{(i)} \cap \Big\{\gamma_{j,l}^{(i)} \le L^{(i)}(K-i)\ln(w_{i,t})\Big\}\Big)\\
&\le \mathbb{P}\Big(\Big\{\exists t' \in \mathcal{I}_{j,l}^{(i)}, a \in \mathcal{S}_{l-1}^{(i)} : \theta_{a,t'} > \mu_a + \delta_{i,a}\Big\}\\
&\qquad\qquad\qquad\qquad \cap E_j^{(i)} \cap F_{j,l-1}^{(i)}\Big)\\
&\quad + \underbrace{\mathbb{P}\Big(\Big\{\forall t' \in \mathcal{I}_{j,l}^{(i)}, a \in \mathcal{S}_{l-1}^{(i)} : \theta_{a,t'} \le \mu_a + \delta_{i,a}\Big\} \cap E_j^{(i)} \cap F_{j,l-1}^{(i)} \cap \Big\{\gamma_{j,l}^{(i)} \le L^{(i)}(K-i)\ln(w_{i,t})\Big\}\Big)}_{B}\\
&\le \mathbb{P}\Big(\exists t' \in \mathcal{I}_{j,l}^{(i)}, a > i : \theta_{a,t'} > \mu_a + \delta_{i,a},\\
&\qquad\qquad n_{a,t'}^{(i)} > L_{i,a}\ln(w_{i,t})\Big) + \mathbb{P}(B)\\
&\le \frac{2}{w_{i,t}^{2+b}} + \mathbb{P}(B). \qquad \text{(Using Lemma 5)}
\end{aligned}$$

In order to bound $\mathbb{P}(B)$, we define for $k \in \{0,\dots,\gamma_{j,l}^{(i)}\}$, the random intervals $\mathcal{J}_k^{(i)}$ as the time range between k^{th} and $(k+1)^{\text{th}}$ interruption in $\mathcal{I}_{j,l}^{(i)}$. The event B implies there is an interval of $\mathcal{I}_{j,l}^{(i)}$ of length $\left\lceil \frac{w_{i,t}^{1-b}-1}{L^{(i)}(K-i)^2\ln(w_{i,t})} \right\rceil$ during which there are no interruptions, which, in turn, implies that the Thompson samples of the unsaturated arms are smaller than the highest Thompson sample of the saturated arms. Thus,

$$\begin{aligned}
&\mathbb{P}(B)\\
&\le \mathbb{P}\Bigg(\Bigg\{\exists k \in \{0,\dots,\gamma_{j,l}^{(i)}\} : |\mathcal{J}_k^{(i)}| \ge \frac{w_{i,t}^{1-b}-1}{L^{(i)}(K-i)^2\ln(w_{i,t})}\Bigg\}\\
&\qquad \cap \{\forall t' \in \mathcal{I}_{j,l}^{(i)}, a \in \mathcal{S}_{l-1}^{(i)} : \theta_{a,t'} \le y_i\} \cap E_j^{(i)} \cap F_{j,l-1}^{(i)}\Bigg)\\
&\qquad\qquad \text{(where } y_i = \mu_{i+1} + \delta_{i,i+1}\text{)}\\
&\le \sum_{k=1}^{(K-i)L^{(i)}\ln(w_{i,t})} \mathbb{P}\Bigg(\Bigg\{|\mathcal{J}_k^{(i)}| \ge \frac{w_{i,t}^{1-b}-1}{L^{(i)}(K-i)^2\ln(w_{i,t})}\Bigg\}\\
&\qquad \cap \{\forall s \in \mathcal{J}_k^{(i)}, a > i : \theta_{a,s} \le y_i\} \cap E_j^{(i)}\Bigg)\\
&\le \sum_{k=1}^{(K-i)L^{(i)}\ln(w_{i,t})} \mathbb{P}\Bigg(\Bigg\{|\mathcal{J}_k^{(i)}| \ge \frac{w_{i,t}^{1-b}-1}{L^{(i)}(K-i)^2\ln(w_{i,t})}\Bigg\}\\
&\qquad \cap \{\theta_{i,s} \le y_i\} \cap E_j^{(i)}\Bigg)
\end{aligned}$$

$$\begin{aligned}
&\le (K-i)L^{(i)}\ln(w_{i,t})(\alpha_i)^{\frac{w_{i,t}^{1-b}-1}{L^{(i)}(K-i)^2\ln(w_{i,t})}}\\
&\quad + C_{\lambda,i}\frac{L^{(i)}(K-i)\ln(w_{i,t})}{\left(\frac{w_{i,t}^{1-b}-1}{L^{(i)}(K-i)^2\ln(w_{i,t})}\right)^{\lambda}} e^{-jd_{\lambda,i}}\\
&:= f(i,b,j,t).
\end{aligned}$$

Once again we get $\displaystyle\sum_{w_{i,t}\ge 1}\sum_{j\le w_{i,t}^b} f(i,j,b,t) < +\infty$.

Putting it all together: From Equation (3) we get

$$\begin{aligned}
\mathbb{P}\Big(E_j^{(i)}\Big) &\le (K-i-1)\left(\frac{2}{w_{i,t}^{2+b}} + f(i,b,j,t)\right)\\
&\quad + \frac{2}{w_{i,t}^{2+b}} + g(i,b,j,t)\\
&\le \frac{2(K-i)}{w_{i,t}^{2+b}} + Kf(i,b,j,t) + g(i,b,j,t).
\end{aligned}$$

Thus, from Equation (2) we obtain

$$\begin{aligned}
&\sum_{t\ge 1}\mathbb{P}\Big(n_{i,t} \le w_{i,t}^b\Big)\\
&\le N_{i,b} + \sum_{w_{i,t}\ge 1}\sum_{j\le a_{i,t}^b}[Kf(i,b,j,t) + g(i,b,j,t)]\\
&\quad + 2(K-i)\sum_{w_{i,t}\ge 1}\sum_{j\le w_{i,t}^b}\frac{1}{w_{i,t}^{2+b}}\\
&\le N_{i,b} + \sum_{w_{i,t}\ge 1}\sum_{j\le w_{i,t}^b}[Kf(i,b,j,t) + g(i,b,j,t)]\\
&\quad + 2(K-i)\sum_{w_{i,t}\ge 1}\frac{1}{w_{i,t}^2}\\
&\le C_{i,b},
\end{aligned}$$

where, $C_{i,b}$ is some finite constant, as desired. □

5 EXPERIMENTAL RESULTS

In this section, we describe our experimental results comparing the performance of TS-SMAB and AUER algorithms for the SMAB problem. We work with SMAB instances where $K = 10$, $T = 10^5$, $\mu_1 = 0.5$, and $\Delta_{i,i+1} = 0.01$ for all $i \in [K-1]$. We consider the following three settings of the availability sequences of the arms: (1) The *classical MAB* setting, where each arm is available at every time instant, (2) the *changing suboptimal arm* setting, where arm 1 is always available, along with one other suboptimal arm (which changes constantly), i.e., $A_1 = \{1,2\}$, $A_2 = \{1,3\}$,…, $A_K = \{1,K\}$, $A_{K+1} = \{1,2\}$ and so on, and (3) the *random availability* setting, where each arm is (independently) available at every time instant with probability 0.5.

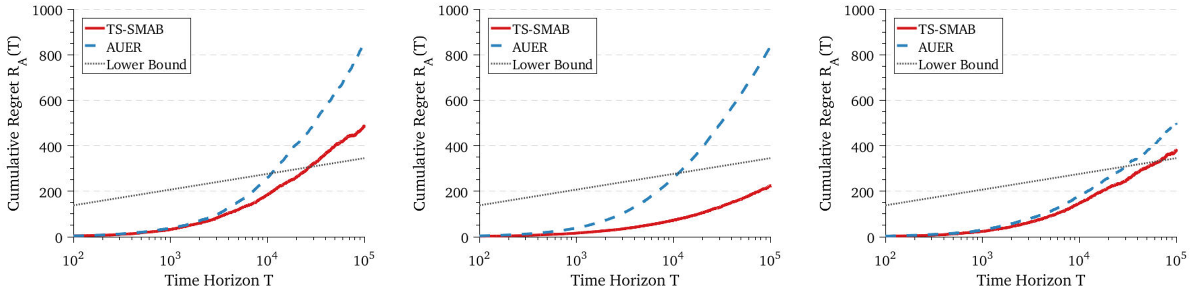

Figure 2: Comparing the regret of AUER and TS-SMAB algorithms for each of the three availability settings.

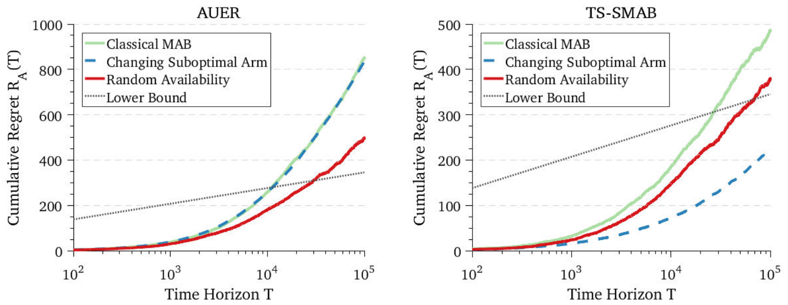

Figure 3: Evaluating each algorithm (AUER and TS-SMAB) across all three availability settings.

Figures 2 and 3 present our results. Each plot in Figure 2 compares the performance of TS-SMAB and AUER for one of the three availability settings described above. Each plot in Figure 3 pertains to a fixed algorithm, and evaluates its performance across the three availability settings. The regret values in each case are averaged over 1000 runs of the algorithm for each value of the time horizon T. We remark that the figures plot the actual expected regret accumulated in the simulations and not the theoretical upper bounds on the regret. We also plot the function $30\ln(T)$ as a proxy for the lower bound for SMAB (refer to Proposition 2).

Figure 2 shows that TS-SMAB outperforms AUER for each of the three availability settings described above, with the contrast being the most pronounced for the "changing suboptimal arm" setting.

Figure 3 shows how each of the two algorithms - AUER and TS-SMAB - performs across a range of availability settings.The performance of TS-SMAB is in agreement with our bound on the availability-specific regret from Theorem 1, which is maximized when $a_{j,T} = T$ for all $j \in [K]$ (as in the classical MAB setting), closely followed by the case when $a_{j,T} \approx T/2$ (as in the random availability setting), and is the smallest when $a_{j,T} = T/(K-1)$ for $j \in \{2, \ldots, K\}$ (as in the changing suboptimal arm setting). By contrast, the regret of AUER for the classical and the 'changing suboptimal arms' setting is nearly indistinguishable.

6 CONCLUDING REMARKS

We studied the sleeping multi-armed bandit problem with stochastic rewards and adversarial availabilities, and showed an $\mathcal{O}(\log T)$ regret bound for the Thompson Sampling algorithm (TS-SMAB). We also showed experimentally that TS-SMAB outperforms the AUER algorithm for various settings of availabilities, and discussed how our bounds capture the dependence of the regret of TS-SMAB on the structure of the availabilities.

As part of future work, it would be interesting to find out whether TS-SMAB is asymptotically optimal. It would also be of general interest to investigate problem independent bounds on the regret. Lastly, we intend to reproduce our experimental results on real world datasets.

Acknowledgments

The work reported here is partially supported by and is a part of the ongoing project *Disaster Recovery using Game Theory and Multiagent Systems* sponsored by the Defence Research and Development Organisation (DRDO), Government of India.

References

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson Sampling for the multi-armed bandit problem. In *25th Annual Conference on Learning Theory, COLT 2012*, pages 39.1–39.26.

Agrawal, S. and Goyal, N. (2013a). Further optimal regret bounds for Thompson Sampling. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics, AISTATS 2013*, pages 99–107.

Agrawal, S. and Goyal, N. (2013b). Thompson Sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, pages 127–135.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning 2002*, 47(2-3):235–256.

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

Chapelle, O. and Li, L. (2011). An empirical evaluation of Thompson Sampling. In *25th Annual Conference on Neural Information Processing Systems (NIPS) 2011*, pages 2249–2257.

Gentile, C., Li, S., and Zappella, G. (2014). Online clustering of bandits. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 757–765.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association, 1963*, 58(301):13–30.

Kanade, V., McMahan, H. B., and Bryan, B. (2009). Sleeping experts and bandits with stochastic action availability and adversarial rewards. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, AISTATS 2009*, pages 272–279.

Kanade, V. and Steinke, T. (2014). Learning hurdles for sleeping experts. *ACM Transactions on Computation Theory, TOCT 2014*, 6(3):11:1–11:16.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson Sampling: An asymptotically optimal finite-time analysis. In *23rd International Conference on Algorithmic Learning Theory, ALT 2012*, pages 199–213.

Kleinberg, R., Niculescu-Mizil, A., and Sharma, Y. (2010). Regret bounds for sleeping experts and bandits. *Machine Learning 2010*, 80(2-3):245–272.

Komiyama, J., Honda, J., and Nakagawa, H. (2015). Optimal regret analysis of Thompson Sampling in stochastic multi-armed bandit problem with multiple plays. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, pages 1152–1161.

Li, L. and Chapelle, O. (2012). Open problem: Regret bounds for thompson sampling. In *25th Annual Conference on Learning Theory, COLT 2012*, pages 43.1–43.3.

Li, S., Karatzoglou, A., and Gentile, C. (2016). Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 539–548. ACM.

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

Xia, Y., Li, H., Qin, T., Yu, N., and Liu, T. (2015). Thompson Sampling for budgeted multi-armed bandits. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence, IJCAI 2015*, pages 3960–3966.