MOTS: Minimax Optimal Thompson Sampling

Tianyuan Jin[1] **Pan Xu**[2] **Jieming Shi**[3] **Xiaokui Xiao**[1] **Quanquan Gu**[2]

Abstract

Thompson sampling is one of the most widely used algorithms for many online decision problems, due to its simplicity in implementation and superior empirical performance over other state-of-the-art methods. Despite its popularity and empirical success, it has remained an open problem whether Thompson sampling can match the minimax lower bound $\Omega(\sqrt{KT})$ for K-armed bandit problems, where T is the total time horizon. In this paper, we solve this long open problem by proposing a variant of Thompson sampling called MOTS that adaptively clips the sampling instance of the chosen arm at each time step. We prove that this simple variant of Thompson sampling achieves the minimax optimal regret bound $O(\sqrt{KT})$ for finite time horizon T, as well as the asymptotic optimal regret bound for Gaussian rewards when T approaches infinity. To our knowledge, MOTS is the first Thompson sampling type algorithm that achieves the minimax optimality for multi-armed bandit problems.

1. Introduction

The Multi-Armed Bandit (MAB) problem is a sequential decision process which is typically described as a game between the agent and the environment with K arms. The game proceeds in T time steps. In each time step $t = 1, \ldots, T$, the agent plays an arm $A_t \in \{1, 2, \cdots, K\}$ based on the observation of the previous $t-1$ time steps, and then observes a reward r_t that is independently generated from a 1-subGaussian distribution with mean value μ_{A_t}, where $\mu_1, \mu_2, \cdots, \mu_K \in \mathbb{R}$ are unknown. The goal of the agent is to maximize the cumulative reward over T time steps. The performance of a strategy for MAB is measured by the expected cumulative difference over T time steps between playing the best arm and playing the arm according to the strategy, which is also called the regret of a bandit strategy. Formally, the regret $R_\mu(T)$ is defined as follows

$$R_\mu(T) = T \cdot \max_{i \in \{1,2,\cdots,K\}} \mu_i - \mathbb{E}_\mu\left[\sum_{t=1}^{T} r_t\right]. \qquad (1)$$

For a fixed time horizon T, the problem-independent lower bound (Auer et al., 2002b) states that any strategy has at least a regret in the order of $\Omega(\sqrt{KT})$, which is called the *minimax optimal* regret. On the other hand, for a fixed model (i.e., $\mu_1, \ldots, \mu_K$ are fixed), Lai & Robbins (1985) proved that any strategy must have at least $C(\mu)\log(T)(1-o(1))$ regret when the horizon T approaches infinity, where $C(\mu)$ is a constant depending on the model. Therefore, a strategy with a regret upper-bounded by $C(\mu)\log(T)(1-o(1))$ is *asymptotically optimal*.

This paper studies the earliest bandit strategy, Thompson sampling (TS) (Thompson, 1933). It has been observed in practice that TS often achieves a smaller regret than many upper confidence bound (UCB)-based algorithms (Chapelle & Li, 2011; Wang & Chen, 2018). In addition, TS is simple and easy to implement. Despite these advantages, the theoretical analysis of TS algorithms has not been established until the past decade. In particular, in the seminal work by Agrawal & Goyal (2012), they provided the first finite-time analysis of TS. Kaufmann et al. (2012) and Agrawal & Goyal (2013) showed that the regret bound of TS is asymptotically optimal when using Beta priors. Subsequently, Agrawal & Goyal (2017) showed that TS with Beta priors achieves an $O(\sqrt{KT\log T})$ problem-independent regret bound while maintaining the asymptotic optimality. In addition, they proved that TS with Gaussian priors can achieve an improved regret bound $O(\sqrt{KT\log K})$. Agrawal & Goyal (2017) also established the following regret lower bound for TS: the TS strategy with Gaussian priors has a worst-case regret $\Omega(\sqrt{KT\log K})$.

Main Contributions. It remains an open problem (Li & Chapelle, 2012) whether TS type algorithms can achieve the minimax optimal regret bound $O(\sqrt{KT})$ for MAB problems. In this paper, we solve this open problem by proposing a variant of Thompson sampling, referred to as Minimax Optimal Thompson Sampling (MOTS), which clips the sampling instances for each arm based on the his-

[1]School of Computing, National University of Singapore, Singapore [2]Department of Computer Science, University of California, Los Angeles, USA [3]Department of Computing, The Hong Kong Polytechnic University, Hong Kong. Correspondence to: Xiaokui Xiao <xkxiao@nus.edu.sg>, Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Table 1. Comparisons of different TS type algorithms. The *minimax ratio* is the ratio (up to constant factors) between the problem-independent regret bound of the algorithm and the minimax optimal regret $O(\sqrt{KT})$. For instance, when the ration equals 1, it is minimax optimal; otherwise, it is minimax suboptimal. The results in Kaufmann et al. (2012); Agrawal & Goyal (2013; 2017) are obtained for rewards bounded in $[0, 1]$, but the techniques in their papers also work for Gaussian rewards (See Korda et al. (2013) for details).

	Reward Type	Minimax Ratio	Asym. Optimal	Reference
TS	Bernoulli	–	Yes	Kaufmann et al. (2012)
	Bernoulli	$\sqrt{\log T}$	Yes	Agrawal & Goyal (2013)
	Bernoulli	$\sqrt{\log K}$ *	–	Agrawal & Goyal (2017)
MOTS	subGaussian	1	No**	▷ Theorems 1, 2
	subGaussian	$\text{ilog}^{(m-1)}(T)$ ***	Yes	▷ Theorem 3
MOTS-$\mathcal{J}$	Gaussian	1	Yes	▷ Theorem 4

* It has been proved by Agrawal & Goyal (2017) that the $\sqrt{\log K}$ term in the problem-independent regret is unimprovable for Thompson sampling using Gaussian priors.

** As is shown in Theorem 2, MOTS is asymptotically optimal up to a multiplicative factor $1/\rho$, where $\rho \in (1/2, 1)$ is a fixed constant.

*** $\text{ilog}^{(r)}(\cdot)$ is the iterated logarithm of order r, and $m \geq 2$ is an arbitrary integer independent of T.

tory of pulls. We prove that MOTS achieves $O(\sqrt{KT})$ problem-independent regret, which is minimax optimal and improves the existing best result, i.e., $O(\sqrt{KT \log K})$. Furthermore, we show that when the reward distributions are Gaussian, a variant of MOTS with clipped Rayleigh distributions, namely MOTS-$\mathcal{J}$, can simultaneously achieve asymptotic and minimax optimal regret bounds. Our result also conveys the important message that the lower bound for TS with Gaussian priors in Agrawal & Goyal (2017) may not always hold in the more general cases when non-Gaussian priors are used. Our experiments demonstrate the superiority of MOTS over the state-of-the-art bandit algorithms such as UCB (Auer et al., 2002a), MOSS (Audibert & Bubeck, 2009), and TS (Thompson, 1933) with Gaussian priors. We provide a detailed comparison on the minimax optimality and asymptotic optimality of TS type algorithms in Table 1.

Notations. A random variable X is said to follow a 1-subGaussian distribution, if it holds that $\mathbb{E}_X[\exp(\lambda X - \lambda \mathbb{E}_X[X])] \leq \exp(\lambda^2/2)$ for all $\lambda \in \mathbb{R}$. We denote $\log^+(x) = \max\{0, \log x\}$. We let T be the total number of time steps, K be the number of arms, and $[K] = \{1, 2, \cdots, K\}$. Without loss of generality, we assume that $\mu_1 = \max_{i \in [K]} \mu_i$ throughout this paper. We use Δ_i to denote the gap between arm 1 and arm i, i.e., $\Delta_i := \mu_1 - \mu_i$, $i \in [K] \setminus \{1\}$. We denote $T_i(t) := \sum_{j=1}^{t} \mathbb{1}\{A_j = i\}$ as the number of times that arm i has been played at time step t, and $\widehat{\mu}_i(t) := \sum_{j=1}^{t} \mathbb{1}\{A_j = i\} \cdot r_j / T_i(t)$ as the average reward for pulling arm i up to time t, where r_j is the reward received by the algorithm at time j.

2. Related Work

Existing works on regret minimization for stochastic bandit problems mainly consider two notions of optimality: asymptotic optimality and minimax optimality. UCB (Garivier & Cappé, 2011; Maillard et al., 2011), Bayes UCB (Kaufmann, 2016), and Thompson sampling (Kaufmann et al., 2012; Agrawal & Goyal, 2017; Korda et al., 2013) are all shown to be asymptotically optimal. Meanwhile, MOSS (Audibert & Bubeck, 2009) is the first method proved to be minimax optimal. Subsequently, two UCB-based methods, AdaUCB (Lattimore, 2018) and KL-UCB^{++} (Ménard & Garivier, 2017), are also shown to achieve minimax optimality. In addition, AdaUCB is proved to be almost instance-dependent optimal for Gaussian multi-armed bandit problems (Lattimore, 2018).

There are many other methods on regret minimization for stochastic bandits, including explore-then-commit (Auer & Ortner, 2010; Perchet et al., 2016), ϵ-Greedy (Auer et al., 2002a), and RandUCB (Vaswani et al., 2019). However, these methods are proved to be suboptimal (Auer et al., 2002a; Garivier et al., 2016; Vaswani et al., 2019). One exception is the recent proposed double explore-then-commit algorithm (Jin et al., 2020), which achieves asymptotic optimality. Another line of works study different variants of the problem setting, such as the batched bandit problem (Gao et al., 2019), and bandit with delayed feedback (Pike-Burke et al., 2018). We refer interested readers to Lattimore & Szepesvári (2020) for a more comprehensive overview of bandit algorithms.

For Thompson sampling, Russo & Van Roy (2014) studied the Bayesian regret and Bubeck & Liu (2013) improved it to $O(\sqrt{KT})$ using the confidence bound analysis of MOSS (Audibert & Bubeck, 2009). However, it should be noted that the Bayesian regret is known to be less informative than the frequentist regret $R_\mu(T)$ studied in this paper. In fact, one can show that our minimax optimal regret $R_\mu(T) = O(\sqrt{KT})$ immediately implies that the Bayesian regret is also in the order of $O(\sqrt{KT})$, but the reverse is not true (Lattimore & Szepesvári, 2020). We refer interested readers to Russo et al. (2018) for a thorough introduction of Thompson sampling and its various applications.

3. Minimax Optimal Thompson Sampling Algorithm

3.1. General Thompson sampling strategy

We first describe the general Thompson sampling (TS) strategy. In the first K time steps, TS plays each arm $i \in [K]$ once, and updates the average reward estimation $\widehat{\mu}_i(K+1)$ for each arm i. (This is a standard warm-start in the bandit literature.) Subsequently, the algorithm maintains a distribution $D_i(t)$ for each arm $i \in [K]$ at time step $t = K+1, \ldots, T$, whose update rule will be elaborated shortly. At step t, the algorithm samples instances $\theta_i(t)$ independently from distribution $D_i(t)$, for all $i \in [K]$. Then, the algorithm plays the arm that maximizes $\theta_i(t)$: $A_t = \mathrm{argmax}_{i \in [K]} \theta_i(t)$, and receives a reward r_t. The average reward $\widehat{\mu}_i(t)$ and the number of pulls $T_i(t)$ for arm $i \in [K]$ are then updated accordingly.

We refer to algorithms that follow the general TS strategy described above (e.g., those in Chapelle & Li (2011); Agrawal & Goyal (2017)) as *TS type algorithms*. Following the above definition, our MOTS method is a TS type algorithm, but it differs from other algorithms of this type in the choice of distribution $D_i(t)$: existing algorithms (e.g., Agrawal & Goyal (2017)) typically use Gaussian or Beta distributions as the posterior distribution, whereas MOTS uses a *clipped* Gaussian distribution, which we detail in Section 3.2. Nevertheless, we should note that MOTS fits exactly into the description of Thompson sampling in Li & Chapelle (2012); Chapelle & Li (2011).

3.2. Thompson sampling using clipped Gaussian distributions

Algorithm 1 shows the pseudo-code of MOTS, with $D_i(t)$ formulated as follows. First, at time step t, for all arm $i \in [K]$, we define a *confidence range* $(-\infty, \tau_i(t))$, where

$$\tau_i(t) = \widehat{\mu}_i(t) + \sqrt{\frac{\alpha}{T_i(t)} \log^+\left(\frac{T}{KT_i(t)}\right)}, \qquad (2)$$

$\log^+(x) = \max\{0, \log x\}$, and $\alpha > 0$ is a constant. Given $\tau_i(t)$, we first sample an instance $\widetilde{\theta}_i(t)$ from Gaussian distribution $\mathcal{N}(\widehat{\mu}_i(t), 1/(\rho T_i(t)))$, where $\rho \in (1/2, 1)$ is a tuning parameter (The intuition could be found at Lemma 5). Then, we set $\theta_i(t)$ in Line 4 of Algorithm 1 as follows:

$$\theta_i(t) = \min\left\{\widetilde{\theta}_i(t),\ \tau_i(t)\right\}. \qquad (3)$$

In other words, $\theta_i(t)$ follows a *clipped* Gaussian distribution with the following PDF:

$$f(x) = \begin{cases} \varphi_t(x) + (1 - \Phi_t(x))\delta(x - \tau_i(t)), & \text{if } x \le \tau_i(t); \\ 0, & \text{otherwise}, \end{cases} \qquad (4)$$

where $\varphi_t(x)$ and $\Phi_t(x)$ denote the PDF and CDF of the Gaussian distribution $\mathcal{N}(\widehat{\mu}_i(t), \frac{1}{\rho T_i(t)})$, respectively, and $\delta(\cdot)$ is the Dirac delta function.

Algorithm 1 Minimax Optimal Thompson Sampling with Clipping (MOTS)

```
1: Input: Arm set $[K]$.
2: Initialization: Play arm once and set $T_i(K+1) = 1$;
   let $\widehat{\mu}_i(K+1)$ be the observed reward of playing arm $i$
3: for $t = K+1, K+2, \cdots, T$ do
4:     For all $i \in [K]$, sample $\theta_i(t)$ independently from
       $D_i(t)$, which is defined in Section 3.2
5:     Play arm $A_t = \arg\max_{i \in [K]} \theta_i(t)$ and observe the
       reward $r_t$
6:     For all $i \in [K]$
           $\widehat{\mu}_i(t+1) = \frac{T_i(t) \cdot \widehat{\mu}_i(t) + r_t \mathbb{1}\{i = A_t\}}{T_i(t) + \mathbb{1}\{i = A_t\}}$
7:     For all $i \in [K]$: $T_i(t+1) = T_i(t) + \mathbb{1}\{i = A_t\}$
8: end for
```

Remark 1. (2) *dependents on the horizon T, which sometimes be unknown. By using the anytime MOSS, i.e., replace T by t in* (2) *(see Degenne & Perchet (2016) for details), one can make MOTS anytime.*

3.3. Overestimation and underestimation in Thompson sampling

Compared with vanilla TS (see Agrawal & Goyal (2017) for example), MOTS is different in the following two aspects: (1) $\theta_i(t)$ in (3) is sampled from a clipped Gaussian distribution instead of the common Gaussian distribution; (2) before the clipping step, $\widetilde{\theta}_i(t)$ is sampled from a Gaussian distribution whose variance is inflated by a factor $1/\rho$. Both the clipping step and the inflation are crucial for improving the regret bound of vanilla TS to be minimax optimal, which address the overestimation of suboptimal arms and the underestimation of the optimal arm in vanilla TS.

Overestimation of suboptimal arms: at any time step t, vanilla TS needs to ensure the posterior sample of the optimal arm $\theta_1(t)$ is larger than that of all $K-1$ suboptimal arms. If the sample from each suboptimal arm has a probability p to exceed $\theta_1(t)$, then by the union bound, the total probability of identifying the wrong arm will be $(K-1)p$, which leads to a $\sqrt{\log K}$ factor in the regret bound. MOTS clips the posterior samples by a carefully chosen threshold for each arm, which avoids the case that suboptimal arms are overestimated to a large extent.

Underestimation of the optimal arm: with the clipping step, vanilla TS will still fail to find the optimal arm if its posterior sample $\theta_1(t)$ is too small. In this case, the clipping threshold of arm 1 will become smaller in the next step, and then it will be further underestimated. This means the underestimation of the optimal arm will cause severe consequence: it will hardly be picked again once underestimated. We refer readers to Lemma 5 and its discussion for more details. In contrast, MOTS enlarges the variance of the posterior by a factor of $\sqrt{1/\rho}$, where $\rho \in (0,1)$. This increases the probability that the posterior sample of arm 1 before clipping is larger or equals to the threshold, i.e., $\mathbb{P}(\widetilde{\theta}_1 \geq \tau_1)$ will become larger.

In Section 4, we will show that with the help of clipping and the inflation, MOTS achieves the minimax optimality.

4. Theoretical Analysis of MOTS

4.1. Regret of MOTS for subGaussian rewards

We first show that MOTS is minimax optimal.

Theorem 1 (Minimax Optimality of MOTS)**.** *Assume that the reward of arm $i \in [K]$ is 1-subGaussian with mean μ_i. For any fixed $\rho \in (1/2, 1)$ and $\alpha \geq 4$, the regret of Algorithm 1 satisfies*

$$R_\mu(T) = O\bigg(\sqrt{KT} + \sum_{i=2}^{K} \Delta_i\bigg). \tag{5}$$

The second term on the right hand side of (5) is due to the fact that we need to pull each arm at least once in Algorithm 1. Following the convention in the literature (Audibert & Bubeck, 2009; Agrawal & Goyal, 2017), we only need to consider the case when $\sum_{i=2}^{K} \Delta_i$ is dominated by $\sqrt{KT}$.

Remark 2. *Compared with the results in Agrawal & Goyal (2017), the regret bound of MOTS improves that of TS with Beta priors by a factor of $O(\sqrt{\log T})$, and that of TS with Gaussian priors by a factor of $O(\sqrt{\log K})$. To the best of our knowledge, MOTS is the first TS type algorithm that achieves the minimax optimal regret $\Omega(\sqrt{KT})$ for MAB problems (Auer et al., 2002a).*

The next theorem presents the asymptotic regret bound of MOTS for subGaussian rewards.

Theorem 2. *Under the same conditions in Theorem 1, the regret $R_\mu(T)$ of Algorithm 1 satisfies*

$$\lim_{T\to\infty} \frac{R_\mu(T)}{\log(T)} = \sum_{i:\Delta_i>0} \frac{2}{\rho\Delta_i}. \tag{6}$$

Lai & Robbins (1985) proved that for Gaussian rewards, the asymptotic regret rate $\lim_{T\to\infty} R_\mu/\log T$ is at least $\sum_{i:\Delta_i>0} 2/\Delta_i$. Therefore, Theorem 2 indicates that the asymptotic regret rate of MOTS matches the aforementioned lower bound up to a multiplicative factor $1/\rho$, where $\rho \in (1/2, 1)$ is arbitrarily fixed.

In the following theorem, by setting ρ to be time-varying, we show that MOTS is able to exactly match the asymptotic lower bound.

Theorem 3. *Assume the reward of each arm i is 1-subGaussian with mean μ_i, $i \in [K]$. In Algorithm 1, if we choose $\alpha \geq 4$ and $\rho = 1 - (\text{ilog}^{(m)}(T)/40)^{-1/2}$, then the regret of MOTS satisfies*

$$R_\mu(T) = O\bigg(\sqrt{KT}\,\text{ilog}^{(m-1)}(T) + \sum_{i=2}^{K} \Delta_i\bigg),$$

$$\lim_{T\to\infty} \frac{R_\mu(T)}{\log(T)} = \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}, \tag{7}$$

where $m \in O_T(1)$ is an arbitrary integer independent of T and $\text{ilog}^{(m)}(T)$ is the result of iteratively applying the logarithm function on T for m times, i.e., $\text{ilog}^{(m)}(x) = \max\big\{\log\big(\text{ilog}^{(m-1)}(x)\big), e\big\}$ and $\text{ilog}^{(0)}(a) = a$.

Theorem 3 indicates that MOTS can exactly match the asymptotic lower bound in Lai & Robbins (1985), at the cost of forgoing minimax optimality by up to a factor of $O(\text{ilog}^{(m-1)}(T))$. For instance, if we choose $m = 4$, then MOTS is minimax optimal up to a factor of $O(\log\log\log T)$. Although this problem-independent bound is slightly worse than that in Theorem 1, it is still a significant improvement over the best known problem-independent bound $O(\sqrt{KT\log T})$ for asymptotically optimal TS type algorithms (Agrawal & Goyal, 2017).

Finally, it should be noted that the lower bound of the asymptotic regret rate $\lim_{T\to\infty} R_\mu/\log T \geq \sum_{i:\Delta_i>0} 2/\Delta_i$ in Lai & Robbins (1985) was established for Gaussian rewards. Since Gaussian is a special case of subGaussian, the lower bound for the Gaussian case is also a valid lower bound for general subGaussian cases. Therefore, MOTS is asymptotically optimal. Similar arguments are widely adopted in the literature (Lattimore & Szepesvári, 2020).

4.2. Regret of MOTS for Gaussian rewards

In this subsection, we present a variant of MOTS, called MOTS-$\mathcal{J}$, which simultaneously achieves the minimax and asymptotic optimality for Gaussian reward distributions.

Algorithm 2 MOTS-$\mathcal{J}$

```
1: Input: Arm set $[K]$.
2: Initialization: Play arm once and set $T_i(K+1) = 1$; let $\widehat{\mu}_i(K+1)$ be the observed reward of playing arm $i$
3: for $t = K+1, K+2, \cdots, T$ do
4:     For all $i \in [K]$, sample $\theta_i(t)$ independently from $D_i(t)$ as follows: sample $\widetilde{\theta}_i(t)$ from $\mathcal{J}(\widehat{\mu}_i(t), 1/T_i(t))$; set $\theta_i(t) = \min\{\widetilde{\theta}_i(t), \tau_i(t)\}$, where $\tau_i(t)$ is defined in (2)
5:     Play arm $A_t = \arg\max_{i\in[K]} \theta_i(t)$ and observe the reward $r_t$
6:     For all $i \in [K]$
           $$\widehat{\mu}_i(t+1) = \frac{T_i(t) \cdot \widehat{\mu}_i(t) + r_t \mathbb{1}\{i = A_t\}}{T_i(t) + \mathbb{1}\{i = A_t\}}$$
7:     For all $i \in [K]$: $T_i(t+1) = T_i(t) + \mathbb{1}\{i = A_t\}$
8: end for
```

Algorithm 2 shows the pseudo-code of MOTS-$\mathcal{J}$. Observe that MOTS-$\mathcal{J}$ is identical to MOTS, except that in Line 4 of MOTS-$\mathcal{J}$, it samples $\widetilde{\theta}_i(t)$ from a distribution $\mathcal{J}(\widehat{\mu}_i(t), 1/T_i(t))$ instead of the Gaussian distribution used in Section 3.2 for MOTS. The distribution $\mathcal{J}(\mu, \sigma^2)$ has the following PDF:

$$\phi_{\mathcal{J}}(x) = \frac{1}{2\sigma^2} \cdot |x - \mu| \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]. \tag{8}$$

Note that $\mathcal{J}$ is a Rayleigh distribution if it is restricted to $x \ge 0$.

The following theorem shows the minimax and asymptotic optimality of MOTS-$\mathcal{J}$ for Gaussian rewards.

Theorem 4. *Assume that the reward of each arm i follows a Gaussian distribution $\mathcal{N}(\mu_i, 1)$, and that $\alpha \ge 2$ in (2). The regret of MOTS-$\mathcal{J}$ satisfies*

$$R_\mu(T) = O\left(\sqrt{KT} + \sum_{i=2}^{K} \Delta_i\right),$$
$$\lim_{T\to\infty} \frac{R_\mu(T)}{\log(T)} = \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}. \tag{9}$$

Remark 3. *To our knowledge, MOTS-$\mathcal{J}$ is the first TS type algorithm that simultaneously achieves the minimax and asymptotic optimality. Though the clipping threshold of MOTS-$\mathcal{J}$ in (2) looks like the MOSS index in Audibert & Bubeck (2009), there are some key differences in the choice of α, the theoretical analysis and the result. Specifically, Audibert & Bubeck (2009) proved that MOSS with the exploration index $\alpha = 4$ achieves minimax optimality for MAB. It remained an open problem how to improve MOSS to be both minimax and asymptotically optimal until Ménard & Garivier (2017) proposed the KL-UCB^{++} algorithm for exponential families of distributions which implies that MOSS with exploration index $\alpha = 2$ could lead to the asymptotic optimal regret for Gaussian rewards. For more details on the choice of α in MOSS, we refer interested readers to the discussion in Chapter 9.3 of Lattimore & Szepesvári (2020).*

Compared with MOSS index based UCB algorithms, our proposed MOTS-$\mathcal{J}$ is both minimax and asymptotically optimal as long as $\alpha \ge 2$. This flexibility is due to the fact that our theoretical analysis (asymptotic optimal part) based on Thompson sampling is quite different from those based on UCB in Audibert & Bubeck (2009); Ménard & Garivier (2017). Not confined by the choice of the exploration index α, it will be more suitable to design better algorithms based on MOTS-$\mathcal{J}$, e.g., achieving instance-dependent optimality (see Lattimore (2015) for details) while keeping the asymptotic optimality.

5. Proof of the Minimax Optimality of MOTS

In what follows, we prove our main result in Theorem 1, and we defer the proofs of all other results to the appendix. We first present several useful lemmas. Lemmas 1 and 2 characterise the concentration properties of subGaussian random variables.

Lemma 1 (Lemma 9.3 in Lattimore & Szepesvári (2020))**.** *Let $X_1, X_2, \cdots$ be independent and 1-subGaussian random variables with zero means. Denote $\widehat{\mu}_t = 1/t \sum_{s=1}^{t} X_s$. Then, for $\alpha \ge 4$ and any $\Delta > 0$,*

$$\mathbb{P}\left(\exists\, s \in [T] : \widehat{\mu}_s + \sqrt{\frac{\alpha}{s}\log^+\left(\frac{T}{sK}\right)} + \Delta \le 0\right) \le \frac{15K}{T\Delta^2}. \tag{10}$$

Lemma 2. *Let $\omega > 0$ be a constant and $X_1, X_2, \ldots, X_n$ be independent and 1-subGaussian random variables with zero means. Denote $\widehat{\mu}_n = 1/n \sum_{s=1}^{n} X_s$. Then, for $\alpha > 0$ and any $N \le T$,*

$$\sum_{n=1}^{T} \mathbb{P}\left(\widehat{\mu}_n + \sqrt{\frac{\alpha}{n}\log^+\left(\frac{N}{n}\right)} \ge \omega\right)$$
$$\le 1 + \frac{\alpha \log^+(N\omega^2)}{\omega^2} + \frac{3}{\omega^2} + \frac{\sqrt{2\alpha\pi \log^+(N\omega^2)}}{\omega^2}. \tag{11}$$

Next, we introduce a few notations for ease of exposition. Recall that we have defined $\widehat{\mu}_i(t)$ to be the average reward for arm i up to a time t. Now, let $\widehat{\mu}_{is}$ be the average reward for arm i up to when it is played the s-th time. In addition, similar to the definitions of $D_i(t)$ and $\theta_i(t)$, we define D_{is} as the distribution of arm i when it is played the s-th time, and θ_{is} as a sample from distribution D_{is}.

The following lemma upper bounds the expected total number of pulls of each arm. We note that it is first proved by Agrawal & Goyal (2017); here, we use an improved version presented in Lattimore & Szepesvári (2020)[1].

Lemma 3 (Theorem 36.2 in Lattimore & Szepesvári (2020))**.** *Let $\epsilon \in \mathbb{R}^+$. Then, the expected number of times that Algorithm 1 plays arm i is bounded by*

$$\begin{aligned}
&\mathbb{E}[T_i(T)]\\
&= \mathbb{E}\Bigg[\sum_{t=1}^{T} \mathbb{1}\{A_t = i, E_i(t)\}\Bigg] + \mathbb{E}\Bigg[\sum_{t=1}^{T} \mathbb{1}\{A_t = i, E_i^c(t)\}\Bigg]\\
&\le 1 + \mathbb{E}\Bigg[\sum_{s=1}^{T-1}\bigg(\frac{1}{G_{1s}(\epsilon)} - 1\bigg)\Bigg] + \mathbb{E}\Bigg[\sum_{t=K+1}^{T-1} \mathbb{1}\{A_t = i, E_i^c(t)\}\Bigg] \qquad (12)\\
&\le 2 + \mathbb{E}\Bigg[\sum_{s=1}^{T-1}\bigg(\frac{1}{G_{1s}(\epsilon)} - 1\bigg)\Bigg] + \mathbb{E}\Bigg[\sum_{s=1}^{T-1} \mathbb{1}\{G_{is}(\epsilon) > 1/T\}\Bigg], \qquad (13)
\end{aligned}$$

where $G_{is}(\epsilon) = 1 - F_{is}(\mu_1 - \epsilon)$, F_{is} is the CDF of D_{is}, and $E_i(t) = \{\theta_i(t) \le \mu_1 - \epsilon\}$.

Based on the decomposition of (12), one can easily prove the problem-independent regret bound of Thompson Sampling by setting $\epsilon = \Delta_i/2$ and summing up over $i = 1, \ldots, K$ (Agrawal & Goyal, 2017). Similar techniques are also used in proving the regret bound of UCB algorithms (Lattimore & Szepesvári, 2020).

Note that by the definition of D_{is}, $G_{is}(\epsilon)$ is a random variable depending on $\widehat{\mu}_{is}$. For brevity, however, we do not explicitly indicate this dependency by writing $G_{is}(\epsilon)$ as $G_{is}(\epsilon, \widehat{\mu}_{is})$; such shortened notations are also used in Agrawal & Goyal (2017); Lattimore & Szepesvári (2020).

Though $G_{is}(\epsilon)$ is defined based on the clipped Gaussian distribution D_{is}, the right-hand side of (12) and (13) can be bounded in the same way for Gaussian distributions like in Agrawal & Goyal (2017). We need some notations. Let F'_{is} be the CDF of Gaussian distribution $\mathcal{N}(\widehat{\mu}_{is}, 1/(\rho s))$ for any $s \ge 1$. Let $G'_{is}(\epsilon) = 1 - F'_{is}(\mu_1 - \epsilon)$. We have the following lemma.

Lemma 4. *Let $\rho \in (1/2, 1)$ be a constant. Under the conditions in Theorem 1, for any $\epsilon > 0$, there exists a universal constant $c > 0$ such that:*

$$\mathbb{E}\Bigg[\sum_{s=1}^{T-1}\bigg(\frac{1}{G'_{1s}(\epsilon)} - 1\bigg)\Bigg] \le \frac{c}{\epsilon^2}. \qquad (14)$$

Similar quantities are also bounded in Agrawal & Goyal (2017); Lattimore & Szepesvári (2020), which are essential for proving the near optimal problem-independent regret bound for Thompson sampling. However, the upper bound in Lemma 4 is sharper than that in previous papers due to the scaling parameter ρ we choose in our MOTS algorithm. In fact, the requirement $\rho \in (1/2, 1)$ is necessary to obtain such an improved upper bound. In the next lemma, we will show that if we choose $\rho = 1$ as is done in existing work, the second term in the right-hand side of (12) will have a nontrivial lower bound.

Lemma 5. *Assume $K \log T \le \sqrt{T}$. If we set $\rho = 1$, then there exists a bandit instance with $\Delta_i = 2\sqrt{K \log T/T}$ for all $i \in \{2, \cdots, K\}$ such that*

$$\mathbb{E}\Bigg[\frac{1}{G'_{1s}(\epsilon)} - 1\Bigg] \ge \frac{e^{-\frac{s\epsilon^2}{2}}}{s\epsilon^2}, \forall \epsilon > 0, \qquad (15)$$

and the decomposition in (12) will lead to

$$K\Delta_i \cdot \mathbb{E}\Bigg[\sum_{s=1}^{T-1}\bigg(\frac{1}{G'_{1s}(\Delta_i/2)} - 1\bigg)\Bigg] = \Omega(\sqrt{KT \log T}).$$

The above lemma shows that if we set $\rho = 1$, the decomposition in (12) will lead to an unavoidable $\Omega(\sqrt{KT \log T})$ problem-independent regret. Combined with Lemma 4, it indicates that our choice of $\rho \in (1/2, 1)$ in MOTS is crucial to improve the previous analysis and obtain a better regret bound. When the reward distribution is Bernoulli, it is worth noting that Agrawal & Goyal (2017) achieved an improved regret $O(\sqrt{KT \log K})$ by using Gaussian priors. Meanwhile, they also proved that this regret bound is unimprovable for Thompson sampling using Gaussian priors, which leaves a gap in achieving the minimax optimal regret $O(\sqrt{KT})$. In the following proof of Theorem 1, we will show that the clipped Gaussian distribution suffices to close this gap and achieve the $O(\sqrt{KT})$ minimax regret. Moreover, in Theorem 4, we will further show that MOTS-$\mathcal{J}$ can achieve the minimax optimal regret by using the Rayleigh distribution and does not need the requirement on the scaling parameter ρ, which is crucial in proving the asymptotic optimality simultaneously.

Now, we prove the minimax optimality of MOTS.

Proof of Theorem 1. Recall that $\widehat{\mu}_{is}$ is the average reward of arm i when it has been played s times. We define Δ as follows:

$$\Delta = \mu_1 - \min_{1 \le s \le T}\left\{\widehat{\mu}_{1s} + \sqrt{\frac{\alpha}{s}\log^+\left(\frac{T}{sK}\right)}\right\}. \qquad (16)$$

The regret of Algorithm 1 can be decomposed as follows.

$$\begin{aligned}
R_\mu(T) &= \sum_{i:\Delta_i > 0} \Delta_i \mathbb{E}[T_i(T)]\\
&\le \mathbb{E}[2T\Delta] + \mathbb{E}\Bigg[\sum_{i:\Delta_i > 2\Delta} \Delta_i T_i(T)\Bigg]
\end{aligned}$$

[1] Since MOTS plays every arm once at the beginning, (12) starts with $t = K + 1$ and $s = 1$.

$$\le \mathbb{E}[2T\Delta] + 8\sqrt{KT} + \mathbb{E}\bigg[\sum_{i\in S}\Delta_i T_i(T)\bigg], \quad (17)$$

where $S = \{i : \Delta_i > \max\{2\Delta, 8\sqrt{K/T}\}\}$ is an index set. The first term in (17) can be bounded as:

$$\begin{aligned}\mathbb{E}[2T\Delta] &= 2T\int_0^\infty \mathbb{P}(\Delta \ge x)\mathrm{d}x\\ &\le 2T\int_0^\infty \min\left\{1, \frac{15K}{Tx^2}\right\}\mathrm{d}x = 4\sqrt{15KT}, \quad (18)\end{aligned}$$

where the inequality comes from Lemma 1 since

$$\begin{aligned}&\mathbb{P}\left(\mu_1 - \min_{1\le s\le T}\left\{\widehat{\mu}_{1s} + \sqrt{\frac{\alpha}{s}\log^+\left(\frac{T}{sK}\right)}\right\} \ge x\right)\\ =&\mathbb{P}\left(\exists 1 \le s \le T : \mu_1 - \widehat{\mu}_{1s} - \sqrt{\frac{\alpha}{s}\log^+\left(\frac{T}{sK}\right)} - x \ge 0\right).\end{aligned}$$

Now we focus on term $\sum_{i\in S}\Delta_i T_i(T)$. Note that the update rules of Algorithm 1 ensure $D_i(t+1) = D_i(t)$ $(t \ge K+1)$ whenever $A_t \ne i$. We define

$$\tau_{is} = \widehat{\mu}_{is} + \sqrt{\frac{\alpha}{s}\log^+\left(\frac{T}{sK}\right)}. \quad (19)$$

By the definition in (2), we have $\tau_{is} = \tau_i(t)$ when $T_i(t) = s$. From the definition of Δ in (16), for $i \in S$, we have

$$\tau_{1s} = \widehat{\mu}_{1s} + \sqrt{\frac{\alpha}{s}\log^+\left(\frac{T}{sK}\right)} \ge \mu_1 - \Delta \ge \mu_1 - \frac{\Delta_i}{2}. \quad (20)$$

Recall the definition of D_{1s}. Let θ_{1s} be a sample from the clipped distribution D_{1s}. As mentioned in Section 3.2, we obtain θ_{1s} with the following procedure. We first sample $\widetilde{\theta}_{1s}$ from distribution $\mathcal{N}(\widehat{\mu}_{1s}, 1/(\rho s))$. If $\widetilde{\theta}_{1s} < \tau_{1s}$, we set $\theta_{1s} = \widetilde{\theta}_{1s}$; otherwise, we set $\theta_{1s} = \tau_{1s}$. (20) implies that $\mu_1 - \Delta_i/2 \le \tau_{1s}$, where τ_{1s} is the boundary for clipping. Therefore, $\mathbb{P}(\widetilde{\theta}_{1s} \ge \mu_1 - \Delta_i/2) = \mathbb{P}(\theta_{1s} \ge \mu_1 - \Delta_i/2)$. By definition, F'_{is} is the CDF of $\mathcal{N}(\widehat{\mu}_{is}, 1/(\rho s))$ and $G'_{is}(\epsilon) = 1 - F'_{is}(\mu_1 - \epsilon)$. Therefore, for any $i \in S$, $G_{1s}(\Delta_i/2) = \mathbb{P}(\theta_{1s} \ge \mu_1 - \Delta_i/2) = \mathbb{P}(\widetilde{\theta}_{1s} \ge \mu_1 - \Delta_i/2) = G'_{1s}(\Delta_i/2)$.

Using (12) of Lemma 3 and setting $\epsilon = \Delta_i/2$, for any $i \in S$, we have

$$\begin{aligned}\Delta_i\mathbb{E}[T_i(T)] &\le \Delta_i + \Delta_i \cdot \mathbb{E}\bigg[\sum_{t=K+1}^{T-1}\mathbb{1}\{A_t = i, E_i^c(t)\}\bigg]\\ &\quad + \Delta_i \cdot \mathbb{E}\bigg[\sum_{s=1}^{T-1}\left(\frac{1}{G_{1s}(\Delta_i/2)} - 1\right)\bigg]\\ &= \Delta_i + \underbrace{\Delta_i \cdot \mathbb{E}\bigg[\sum_{t=K+1}^{T-1}\mathbb{1}\{A_t = i, E_i^c(t)\}\bigg]}_{I_1}\\ &\quad + \underbrace{\Delta_i \cdot \mathbb{E}\bigg[\sum_{s=1}^{T-1}\left(\frac{1}{G'_{1s}(\Delta_i/2)} - 1\right)\bigg]}_{I_2}.\end{aligned}$$

Bounding term I_1: Note that

$$\begin{aligned}E_i^c(t) &= \left\{\theta_i(t) > \mu_1 - \frac{\Delta_i}{2}\right\}\\ &\subseteq \left\{\widehat{\mu}_i(t) + \sqrt{\frac{\alpha}{T_i(t)}\log^+\left(\frac{T}{KT_i(t)}\right)} > \mu_1 - \frac{\Delta_i}{2}\right\}.\end{aligned}$$

We define the following notation:

$$\kappa_i = \sum_{s=1}^{T}\mathbb{1}\left\{\widehat{\mu}_{is} + \sqrt{\frac{\alpha}{s}\log^+\left(\frac{T}{sK}\right)} > \mu_1 - \frac{\Delta_i}{2}\right\}, \quad (21)$$

which immediately implies that

$$I_1 = \Delta_i \cdot \mathbb{E}\bigg[\sum_{t=K+1}^{T-1}\mathbb{1}\{A_t = i, E_i^c(t)\}\bigg] \le \Delta_i\mathbb{E}[\kappa_i]. \quad (22)$$

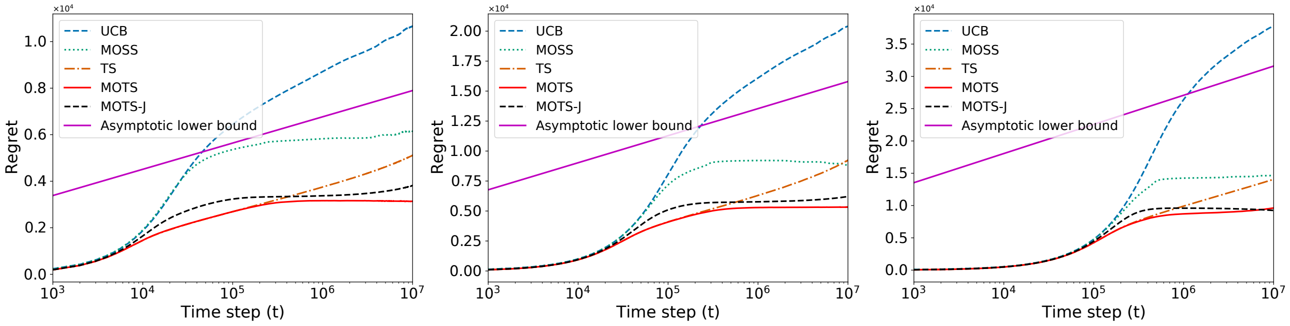
To further bound (22), we have

$$\begin{aligned}&\Delta_i\mathbb{E}[\kappa_i]\\ &= \Delta_i\mathbb{E}\left[\sum_{s=1}^{T}\mathbb{1}\left\{\widehat{\mu}_{is} + \sqrt{\frac{\alpha}{s}\log^+\left(\frac{T}{sK}\right)} > \mu_1 - \frac{\Delta_i}{2}\right\}\right]\\ &\le \Delta_i\sum_{s=1}^{T}\mathbb{P}\left\{\widehat{\mu}_{is} - \mu_i + \sqrt{\frac{\alpha}{s}\log^+\left(\frac{T}{sK}\right)} > \frac{\Delta_i}{2}\right\}\\ &\le \Delta_i + \frac{12}{\Delta_i} + \frac{4\alpha}{\Delta_i}\left(\log^+\left(\frac{T\Delta_i^2}{4K}\right) + \sqrt{2\alpha\pi\log^+\left(\frac{T\Delta_i^2}{4K}\right)}\right),\end{aligned} \quad (23)$$

where the first inequality is due to the fact that $\mu_1 - \mu_i = \Delta_i$ and the second one is by Lemma 2. For $a > 0$, it can be verified that $h(x) = x^{-1}\log^+(ax^2)$ is monotonically decreasing for $x \ge e/\sqrt{a}$. Since $\Delta_i \ge 8\sqrt{K/T} > e/\sqrt{T/(4K)}$, we have $\log(T\Delta_i^2/(4K))/\Delta_i \le \sqrt{T/K}$. Plugging this into (23), we have $\mathbb{E}[\Delta_i\kappa_i] = O(\sqrt{T/K} + \Delta_i)$.

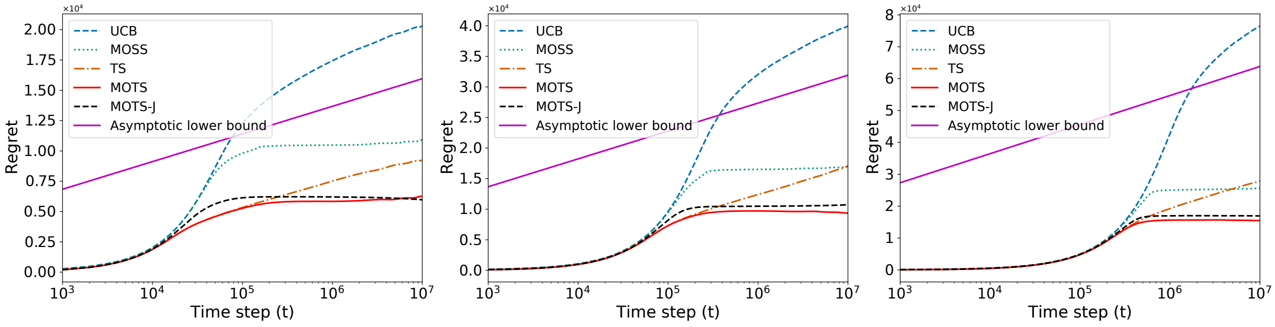
Bounding term I_2: applying Lemma 4, we immediately obtain

$$I_2 = \Delta_i\mathbb{E}\bigg[\sum_{s=1}^{T-1}\left(\frac{1}{G'_{1s}(\Delta_i/2)} - 1\right)\bigg] = O\left(\sqrt{\frac{T}{K}}\right). \quad (24)$$



(a) $K = 50, \epsilon = 0.2$ (b) $K = 50, \epsilon = 0.1$ (c) $K = 50, \epsilon = 0.05$

Figure 1. The regret for different algorithms in Setting (1): $K = 50$ and $\epsilon \in \{0.2, 0.1, 0.05\}$. The experiments are averaged over 6000 repetitions.



(a) $K = 100, \epsilon = 0.2$ (b) $K = 100, \epsilon = 0.1$ (c) $K = 100, \epsilon = 0.05$

Figure 2. The regret for different algorithms in Setting (2): $K = 100$ and $\epsilon \in \{0.2, 0.1, 0.05\}$. The experiments are averaged over 6000 repetitions.

Substituting (18), (21), (23), and (24) into (17), we complete the proof of Theorem 1. □

6. Experiments

In this section, we experimentally compare our proposed algorithms MOTS and MOTS-$\mathcal{J}$ with existing algorithms for multi-armed bandit problems with Gaussian rewards. Baseline algorithms include MOSS (Audibert & Bubeck, 2009), UCB (Katehakis & Robbins, 1995), and Thompson sampling with Gaussian priors (TS for short) (Agrawal & Goyal, 2017). Throughout the experiment, we assume the reward follows an independent Gaussian distribution $\mathcal{N}(\mu, 1)$ with unit variance and mean μ which is to be specified in different settings. Specifically, we consider three settings with different number of arms K and different mean rewards $\{\mu_i\}_{i=1}^K$: (1) $K = 50$, $\mu_1 = 1$, and $\mu_2 = \ldots = \mu_{50} = 1 - \epsilon$, where ϵ varies in the range $\{0.2, 0.1, 0.05\}$ for different experiments; (2) $K = 100$, $\mu_1 = 1$, and $\mu_2 = \ldots = \mu_{100} = 1 - \epsilon$, where ϵ varies in the range $\{0.2, 0.1, 0.05\}$ for different experiments; (3) $K = 50$, $\mu_1 = 1$, $\mu_{5i+j} = 1 - 0.1i$, for $i = 1, \ldots, 9$ and $j = -3, -2, -1, 0, 1$, and $\mu_{47} = \mu_{48} = \mu_{49} = \mu_{50} = 0$. It is worth noting that Setting (1) and Setting (2) only differs in the number of suboptimal arms, where all the suboptimal arms have the same mean value that is distinct from the mean value of the best arm. In contrast, Setting (3) is a more challenging bandit instance where the suboptimal arms are from a rather diverse set.

In all the experiments, the total number of time steps T is set to 10^7. The parameter ρ for MOTS defined in Section 3.2 is set to 0.9999. Since we focus on Gaussian rewards, we set $\alpha = 2$ in (2) for both MOTS and MOTS-$\mathcal{J}$.

Furthermore, for MOTS-$\mathcal{J}$, we need to sample instances from distribution $\mathcal{J}(\mu, \sigma^2)$, of which the PDF is defined in (8). To this end, we use the well known inverse transform sampling technique by first computing the corresponding inverse CDF, and then uniformly choosing a random number in $[0, 1]$, which is then used to calculate the random number sampled from $\mathcal{J}(\mu, \sigma^2)$.

For Setting (1), Figures 1(a), 1(b), and 1(c) report the regrets of all algorithms when ϵ is 0.2, 0.1, 0.05 respectively. For all ϵ values, MOTS consistently outperforms the baselines for all time step t, and MOTS-$\mathcal{J}$ outperforms the baselines especially when t is large. For instance, in Figure 1(c), when time step t is $T = 10^7$, the regret of MOTS and MOTS-$\mathcal{J}$

are 9615 and 9245 respectively, while the regrets of TS, MOSS, and UCB are 14058, 14721, and 37781 respectively. The pink solid line represents the asymptotic lower bound, i.e., $R_\mu(T) = \log(T) \times \sum_{i:\Delta_i>0} 2/\Delta_i$. All the experimental results indicate that our algorithms are asymptotically optimal.

For Setting (2), Figures 2(a), 2(b), and 2(c) report the regrets of MOTS, MOTS-$\mathcal{J}$, MOSS, TS, and UCB when ϵ is $0.2, 0.1, 0.05$ respectively. Again, for all ϵ values, when varying the time step t, MOTS consistently has the smallest regret, outperforming all baselines, and MOTS-$\mathcal{J}$ outperforms all baselines especially when t is large.

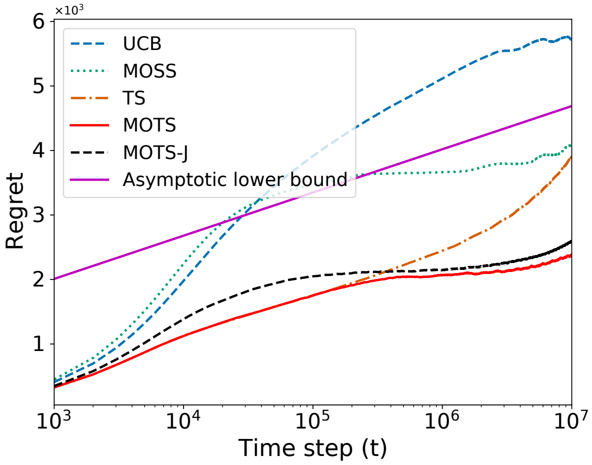


Figure 3. The regret for different algorithms in Setting (3): $K = 50$ and the mean values of arms are from a diverse set. The experiments are averaged over 6000 repetitions.

The experimental results for Setting (3) are reported in Figure 3, which again are consistent with the theoretical findings that MOTS and MOTS-$\mathcal{J}$ outperform baseline algorithms, though the suboptimal arms are extremely diverse and thus the bandit instance is relatively hard. In addition, our algorithms are still asymptotically optimal.

In summary, our algorithms consistently outperform TS, MOSS, and UCB in various settings.

7. Conclusion and Future Work

We solved the open problem on the minimax optimality for Thompson sampling (Li & Chapelle, 2012). We proposed the MOTS algorithm and proved that it achieves the minimax optimal regret $O(\sqrt{KT})$ when rewards are generated from subGaussian distributions. In addition, we propose a variant of MOTS called MOTS-$\mathcal{J}$ that simultaneously achieves the minimax and asymptotically optimal regret for K-armed bandit problems when rewards are generated from Gaussian distributions. Our experiments demonstrate the superior performances of MOTS and MOTS-$\mathcal{J}$ compared with the state-of-the-art bandit algorithms. Interestingly, our experimental results show that the performance of MOTS is never worse than that of MOTS-$\mathcal{J}$. Therefore, it would be an interesting future direction to investigate whether the proposed MOTS with clipped Gaussian distributions can also achieve both minimax and asymptotic optimality for multi-armed bandits.

Acknowledgement

We thank the anonymous reviewers for their helpful comments. X. Xiao is supported by the Ministry of Education, Singapore, under Tier-2 Grant R-252-000-A70-112. T. Jin is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-PhD/2021-01-004[T]). P. Xu and Q. Gu are partially supported by the National Science Foundation IIS-1904183. J. Shi is supported by the financial support (1-BE3T) of research project (P0033898) from Hong Kong Polytechnic University. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Abramowitz, M. and Stegun, I. A. Handbook of mathematical functions with formulas, graphs, and mathematical table. In *US Department of Commerce*. National Bureau of Standards Applied Mathematics series 55, 1965.

Agrawal, S. and Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1, 2012.

Agrawal, S. and Goyal, N. Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pp. 99–107, 2013.

Agrawal, S. and Goyal, N. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5): 30, 2017.

Audibert, J.-Y. and Bubeck, S. Minimax policies for adversarial and stochastic bandits. In *COLT*, pp. 217–226, 2009.

Auer, P. and Ortner, R. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

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

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Bubeck, S. and Liu, C.-Y. Prior-free and prior-dependent regret bounds for thompson sampling. In *Advances in Neural Information Processing Systems*, pp. 638–646, 2013.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pp. 2249–2257, 2011.

Degenne, R. and Perchet, V. Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pp. 1587–1595. PMLR, 2016.

Gao, Z., Han, Y., Ren, Z., and Zhou, Z. Batched multi-armed bandits problem. In *Advances in Neural Information Processing Systems*, pp. 501–511, 2019.

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

Garivier, A., Lattimore, T., and Kaufmann, E. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, pp. 784–792, 2016.

Jin, T., Xu, P., Xiao, X., and Gu, Q. Double explore-then-commit: Asymptotic optimality and beyond. *arXiv preprint arXiv:2002.09174*, 2020.

Katehakis, M. N. and Robbins, H. Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92 (19):8584, 1995.

Kaufmann, E. On bayesian index policies for sequential resource allocation. *arXiv preprint arXiv:1601.01190*, 2016.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pp. 199–213. Springer, 2012.

Korda, N., Kaufmann, E., and Munos, R. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in neural information processing systems*, pp. 1448–1456, 2013.

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

Lattimore, T. Optimally confident ucb: Improved regret for finite-armed bandits. *arXiv preprint arXiv:1507.07880*, 2015.

Lattimore, T. Refining the confidence level for optimistic bandit strategies. *The Journal of Machine Learning Research*, 19(1):765–796, 2018.

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

Li, L. and Chapelle, O. Open problem: Regret bounds for thompson sampling. In *Conference on Learning Theory*, pp. 43–1, 2012.

Maillard, O.-A., Munos, R., and Stoltz, G. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Proceedings of the 24th annual Conference On Learning Theory*, pp. 497–514, 2011.

Ménard, P. and Garivier, A. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pp. 223–237, 2017.

Perchet, V., Rigollet, P., Chassang, S., Snowberg, E., et al. Batched bandit problems. *The Annals of Statistics*, 44(2): 660–681, 2016.

Pike-Burke, C., Agrawal, S., Szepesvari, C., and Grunewalder, S. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pp. 4105–4113, 2018.

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

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

Vaswani, S., Mehrabian, A., Durand, A., and Kveton, B. Old dog learns new tricks: Randomized ucb for bandit problems. *arXiv preprint arXiv:1910.04928*, 2019.

Wang, S. and Chen, W. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 5114–5122, 2018.