Lenient Regret for Multi-Armed Bandits

Nadav Merlis[1] **and Shie Mannor**[1,2]
[1]Technion – Institute of Technology, Israel
[2]Nvidia Research, Israel
merlis@campus.technion.ac.il, shie@ee.technion.ac.il

Abstract

We consider the Multi-Armed Bandit (MAB) problem, where an agent sequentially chooses actions and observes rewards for the actions it took. While the majority of algorithms try to minimize the regret, i.e., the cumulative difference between the reward of the best action and the agent's action, this criterion might lead to undesirable results. For example, in large problems, or when the interaction with the environment is brief, finding an optimal arm is infeasible, and regret-minimizing algorithms tend to over-explore. To overcome this issue, algorithms for such settings should instead focus on playing near-optimal arms. To this end, we suggest a new, more lenient, regret criterion that ignores suboptimality gaps smaller than some ϵ. We then present a variant of the Thompson Sampling (TS) algorithm, called ϵ-TS, and prove its asymptotic optimality in terms of the lenient regret. Importantly, we show that when the mean of the optimal arm is high enough, the lenient regret of ϵ-TS is bounded by a constant. Finally, we show that ϵ-TS can be applied to improve the performance when the agent knows a lower bound of the suboptimality gaps.

Introduction

Multi-Armed Bandit (MAB) problems are sequential decision-making problems where an agent repeatedly chooses an action ('arm'), out of K possible actions, and observes a reward for the selected action (Robbins 1952). In this setting, the agent usually aims to maximize the expected cumulative return throughout the interaction with the problem. Equivalently, it tries to minimize its regret, which is the expected difference between the best achievable total reward and the agent's actual returns.

Although regret is the most prevalent performance criterion, many problems that should intuitively be 'easy' suffer from both large regret and undesired behavior of regret-minimizing algorithms. Consider, for example, a problem where most arms are near-optimal and the few remaining ones have extremely lower rewards. For most practical applications, it suffices to play any of the near-optimal arms, and identifying such arms should be fairly easy. However, regret-minimizing algorithms only compare themselves to the optimal arm. Thus, they must identify an optimal arm with high certainty, or they will suffer linear regret. This leads to two undesired outcomes: (i) the regret fails to characterize the difficulty of such problems, and (ii) regret-minimizing algorithms tend to over-explore suboptimal arms.

Regret fails as a complexity measure: It is well known that for any reasonable algorithm, the regret dramatically increases as the suboptimality gaps shrink, i.e., the reward of some suboptimal arms is very close to the reward of an optimal one (Lai and Robbins 1985). Specifically in our example, if most arms are almost-optimal, then the regret can be arbitrarily large. In contrast, finding a near-optimal solution in this problem is relatively simple. Thus, the regret falsely classifies this easy problem as a hard one.

Regret-minimizing algorithms over-explore: As previously stated, any regret-minimizing agent must identify an optimal arm with high certainty or suffer a linear regret. To do so, the agent must thoroughly explore all suboptimal arms. In contrast, if playing near-optimal arms is adequate, identifying one such arm can be done much more efficiently. Importantly, this issue becomes much more severe in large problems or when the interaction with the problem is brief.

The origin of both problems is the comparison of the agent's reward to the optimal reward. Nonetheless, not all bandit algorithms rely on such comparisons. Notably, when trying to identify good arms ('best-arm identification'), many algorithms only attempt to output ϵ-optimal arms, for some predetermined error level $\epsilon > 0$ (Even-Dar, Mannor, and Mansour 2002). However, this criterion only assesses the quality of the output arms and is unfit when we want the algorithm to choose near-optimal arms throughout the interaction.

In this work, we suggest bringing the leniency of the ϵ-best-arm identification into regret criteria. Inspired by the ϵ-optimality relaxation in best-arm identification, we define the notion of *lenient regret*, that only penalizes arms with gaps larger than ϵ. Intuitively, ignoring small gaps alleviates both previously-mentioned problems: first, arms with gaps smaller than ϵ do not incur lenient regret, and if all other arms have extremely larger gaps, then the lenient regret is expected to be small. Second, removing the penalty from near-optimal arms allows algorithms to spend less time on exploration of bad arms. Then, we expect that algorithms will spend more time playing near-optimal arms.

From a practical perspective, optimizing a more lenient criterion is especially relevant when near-optimal solutions are sufficient while playing bad arms is costly. Consider, for example, a restaurant-recommendation problem. For most people, restaurants of similar quality are practically the same. On the other hand, the cost of visiting bad restaurants is very high. Then, a more lenient criterion should allow focusing on avoiding the bad restaurants, while still recommending restaurants of similar quality.

In the following sections, we formally define the lenient regret and prove a lower bound for this criterion that dramatically improves the classical lower bound (Lai and Robbins 1985) as ϵ increases. Then, inspired by the form of the lower bound, we suggest a variant of the Thompson Sampling (TS) algorithm (Thompson 1933), called ϵ-TS, and prove that its regret asymptotically matches the lower bound, up to an absolute constant. Importantly, we prove that when the mean of the optimal arm is high enough, the lenient regret of ϵ-TS is bounded by a constant. We also provide an empirical evaluation that demonstrates the improvement in performance of ϵ-TS, in comparison to the vanilla TS. Lastly, to demonstrate the generality of our framework, we also show that our algorithm can be applied when the agent has access to a lower bound of all suboptimality gaps. In this case, ϵ-TS greatly improves the performance even in terms of the standard regret.

Related Work

For a comprehensive review of the MAB literature, we refer the readers to (Bubeck, Cesa-Bianchi et al. 2012; Lattimore and Szepesvári 2020; Slivkins et al. 2019). MAB algorithms usually focus on two objectives: regret minimization (Auer, Cesa-Bianchi, and Fischer 2002; Garivier and Cappé 2011; Kaufmann, Korda, and Munos 2012) and best-arm identification (Even-Dar, Mannor, and Mansour 2002; Mannor and Tsitsiklis 2004; Gabillon, Ghavamzadeh, and Lazaric 2012). Intuitively, the lenient regret can be perceived as a weaker regret criterion that borrows the ϵ-optimality relaxation from best-arm identification. Moreover, we will show that in some cases, the lenient regret aims to maximize the number of plays of ϵ-optimal arms. Then, the lenient regret is the most natural adaptation of the ϵ-best-arm identification problem to a regret minimization setting.

Another related concept can be found in sample complexity of Reinforcement Learning (RL) (Kakade 2003; Lattimore et al. 2013; Dann and Brunskill 2015; Dann, Lattimore, and Brunskill 2017). In the episodic setting, this criterion maximizes the number of episodes where an ϵ-optimal policy is played, and can therefore be seen as a possible RL-formulation to our criterion. However, the results for sample complexity significantly differ from ours – first, the lenient regret allows representing more general criteria than the number of ϵ-optimal plays. Second, in the RL settings algorithms focus on the dependence in ϵ and in the size of the state and action spaces, while we derive bounds that depend on the suboptimality gaps. Finally, we show that when the optimal arm is large enough, the lenient regret is constant, and to the best of our knowledge, there is no equivalent result in RL. In some sense, our work can be viewed as a more fundamental analysis of sample complexity that will hopefully allow deriving more general results in RL.

To minimize the lenient regret, we devise a variant of the Thompson Sampling algorithm (Thompson 1933). The vanilla algorithm assumes a prior on the arm distributions, calculates the posterior given the observed rewards and chooses arms according to their probability of being optimal given their posteriors. Even though the algorithm is Bayesian in nature, its regret is asymptotically optimal for any fixed problem (Kaufmann, Korda, and Munos 2012; Agrawal and Goyal 2013a; Korda, Kaufmann, and Munos 2013). The algorithm is known to have superior performance in practice (Chapelle and Li 2011) and has variants for many different settings, i.e., linear bandits (Agrawal and Goyal 2013b), combinatorial bandits (Wang and Chen 2018) and more. For a more detailed review of TS algorithms and their applications, we refer the readers to (Russo et al. 2018). In this work, we present a generalization of the TS algorithm, called ϵ-TS, that minimizes the lenient regret when ignoring gaps smaller than ϵ. Specifically, when $\epsilon = 0$, our approach recovers the vanilla TS.

As previously stated, we also prove that if all gaps are larger than a known $\epsilon > 0$, then our algorithm improves the performance also in terms of the standard regret. Specifically, we prove that the regret of ϵ-TS is bounded by a constant when the optimal arm is larger than $1-\epsilon$. This closely relates to the results of (Bubeck, Perchet, and Rigollet 2013), which proved constant regret bounds when the algorithm knows both the mean of the optimal arm and a lower bound on the gaps. This was later extended in (Lattimore and Munos 2014) for more general structures. Notably, one can apply the results of (Lattimore and Munos 2014) to derive constant regret bounds when all gaps are larger than ϵ and the optimal arm is larger than $1-\epsilon$. Nonetheless, and to the best of our knowledge, we are the first to demonstrate improved performance also when the optimal arm is smaller than $1-\epsilon$.

Setting

We consider the stochastic multi-armed bandit problem with K arms and arm distributions $\underline{\nu} = \{\nu_a\}_{a=1}^K$. At each round, the agent selects an arm $a \in [K] \triangleq \{1, \dots, K\}$. Then, it observes a reward generated from a fixed distribution ν_a, independently at random of other rounds. Specifically, when pulling an arm a on the n^{th} time, it observes a reward $X_{a,n} \sim \nu_a$. We assume that the rewards are bounded in $X_{a,n} \in [0,1]$ and have expectation $\mathbb{E}[X_{a,n}] = \mu_a$. We denote the empirical mean of an arm a using the n first samples by $\hat{\mu}_{a,n} = \frac{1}{n}\sum_{k=1}^n X_{a,k}$ and define $\hat{\mu}_{a,0} = 0$. We also denote the mean of an optimal arm by $\mu^* = \max_a \mu_a$ and the suboptimality gap of an arm a by $\Delta_a = \mu^* - \mu_a$.

Let a_t be the action chosen by the agent at time t. For brevity, we write its gap by $\Delta_t = \Delta_{a_t}$. Next, denote the observed reward after playing a_t by $X_t = X_{a_t, N_{a_t}(t+1)}$, where $N_a(t) = \sum_{\tau=1}^{t-1} \mathbb{1}\{a_\tau = a\}$ is the number of times an arm a was sampled up to time $t-1$. We also let $\hat{\mu}_a(t) = \hat{\mu}_{a,N_a(t)}$, the empirical mean of arm a before round t, and denote the sum over the observed rewards of a up to time $t-1$ by $S_a(t) = \sum_{k=1}^{N_a(t)} X_{a,k} = N_a(t)\hat{\mu}_a(t)$. Finally, we define the natural filtration $\mathcal{F}_t = \sigma(a_1, X_1, \dots, a_t, X_t)$.

Similarly to other TS algorithms, we work with Beta priors. When initialized with parameters $\alpha = \beta = 1$,

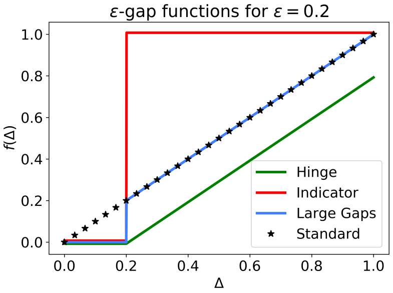

Figure 1: Illustration of different ϵ-gap functions, in comparison to the standard regret $f(\Delta) = \Delta$.

$p \sim \mathrm{Beta}(\alpha, \beta)$ is a uniform distribution. Then, if p is the mean of N Bernoulli experiments, from which there were S 'successes' (ones), the posterior of p is $\mathrm{Beta}(S{+}1, N{-}S{+}1)$. We denote the cumulative distribution function (cdf) of the Beta distribution with parameters $\alpha, \beta > 0$ by $F^{\mathrm{Beta}}_{\alpha,\beta}$. Similarly, we denote the cdf of the Binomial distribution with parameters n, p by $F^{B}_{n,p}$ and its probability density function (pdf) by $f^{B}_{n,p}$. We refer the readers to Appendix A[1] for further details on the distributions and the relations between them (i.e., the 'Beta-Binomial trick'). We also refer the reader to this appendix for some useful concentration results (Hoeffding's inequality and Chernoff-Hoeffding bound).

Finally, we define the Kullback–Leibler (KL) divergence between any two distributions ν and ν' by $\mathrm{KL}(\nu, \nu')$, and let

$$d(p, q) = p \ln \frac{p}{q} + (1 - p) \ln \frac{1 - p}{1 - q} \qquad (1)$$

be the KL-divergence between Bernoulli distributions with means $p, q \in [0, 1]$. By convention, if $p < 1$ and $q \geq 1$, or if $p > 0$ and $q = 0$, we denote $d(p, q) = \infty$.

Regret and Lenient Regret

Most MAB algorithms aim to maximize the expected cumulative reward of the agent. Alternatively, algorithms minimize their expected cumulative regret $R(T) = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_t\right]$. However, and as previously discussed, this sometimes leads to undesired results. Notably, to identify an optimal arm, algorithms must sufficiently explore all suboptimal arms, which is sometimes infeasible. Nonetheless, existing lower bounds for regret-minimizing algorithms show that any reasonable algorithm cannot avoid such exploration (Lai and Robbins 1985). To overcome this issue, we suggest minimizing a weaker notion of regret that ignores small gaps. This will allow finding a near-optimal arm much faster. We formally define this criterion as follows:

Definition 1. *For any $\epsilon \in [0, 1]$, a function $f : [0, 1] \to \mathbb{R}_+$ is called an ϵ-gap function if $f(\Delta) = 0$ for all $\Delta \in [0, \epsilon]$ and $f(\Delta) > 0$ for all $\Delta > \epsilon$. The lenient regret w.r.t. an ϵ-gap function f is defined as $R_f(T) = \mathbb{E}\left[\sum_{t=1}^{T} f(\Delta_t)\right]$.*

While it is natural to require of f to increase with Δ, this assumption is not required for the rest of the paper. Moreover, assuming that $f(\Delta) > 0$ for all $\Delta > \epsilon$ is only required for the lower bound; for the upper bound, it can be replaced by $f(\Delta) \geq 0$ when $\Delta > \epsilon$. There are three notable examples for ϵ-gap functions (see also Figure 1 for graphical illustration). First, the most natural choice for an ϵ-gap function is the hinge loss $f(\Delta) = \max\{\Delta - \epsilon, 0\}$, which ignores small gaps and increases linearly for larger gaps.

Second, we are sometimes interested in maximizing the number of steps where ϵ-optimal arms are played. In this case, we can choose $f(\Delta) = \mathbb{1}\{\Delta > \epsilon\}$. This can be seen as the natural adaptation of ϵ-best-arm identification into a regret criterion. Importantly, notice that this criterion only penalizes sampling of arms with gaps larger than ϵ. This comes with a stark contrast to best-arm identification, where *all* samples are penalized, whether they are of ϵ-optimal arms or not.

Finally, we can choose $f(\Delta) = \Delta \cdot \mathbb{1}\{\Delta > \epsilon\}$. Importantly, when all gaps are larger than ϵ, then this function leads to the standard regret. Thus, all results for ϵ-gap functions also hold for the standard regret when $\Delta_a > \epsilon$ for all suboptimal arms.

There are two ways for relating the lenient regret to the standard regret. First, notice that the standard regret can be represented through the 0-gap function $f(\Delta) = \Delta$. Alternatively, the standard regret can be related to lenient regret w.r.t. the indicator gap-function:

Claim 1. *Let $R(T) = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_t\right]$ be the standard regret and define $f_\epsilon(\Delta) = \mathbb{1}\{\Delta > \epsilon\}$. Then, $R(T) = \int_{\epsilon=0}^{1} R_{f_\epsilon}(T) d\epsilon$.*

The proof is in Appendix E.1. Specifically, it implies that the standard regret aims to minimize the average lenient regret over different leniency levels. In contrast, our approach allows choosing which leniency level to minimize according to the specific application. By doing so, the designer can adjust the algorithm to its needs, instead of using an algorithm that minimizes the average performance.

Lower Bounds

In this section, we prove a problem-dependent lower bound for the lenient regret. Notably, when working with ϵ-gap functions with $\epsilon > 0$, we prove that the lower bound behaves inherently different than the case of $\epsilon = 0$. Namely, for some problems, the lower bound is sub-logarithmic, in contrast to the $\Omega(\ln T)$ bound for the standard regret.

To prove the lower bound, we require some additional notations. Denote by $\mathcal{D}$, a set of distributions over $[0, 1]$ such that $\nu_a \in \mathcal{D}$ for all $a \in [K]$. A bandit strategy is called *consistent* over $\mathcal{D}$ w.r.t. an ϵ-gap function f if for any bandit problem with arm distributions in $\mathcal{D}$ and for any $0 < \alpha \leq 1$, it holds that $R_f(T) = o(T^\alpha)$. Finally, we use $\mathcal{K}_{\inf}$, as was defined in (Burnetas and Katehakis 1996; Garivier, Ménard, and Stoltz 2019):

$$\mathcal{K}_{\inf}(\nu, x, \mathcal{D}) = \inf\{\mathrm{KL}(\nu, \nu') : \nu' \in \mathcal{D}, \mathbb{E}[\nu'] > x\} \ ,$$

and by convention, the infimum over an empty set equals ∞. We now state the lower bound:

Theorem 1. *For any consistent bandit strategy w.r.t. an ϵ-gap function f, for all arms $k \in [K]$ such that $\Delta_k > \epsilon$, it*

[1] A full version can be found at http://arxiv.org/abs/2008.03959.

holds that

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_k(T+1)]}{\ln T} \geq \frac{1}{\mathcal{K}_{\inf}(\nu_k, \mu^* + \epsilon, \mathcal{D})} \ . \tag{2}$$

Specifically, the lenient regret w.r.t. f is lower bounded by

$$\liminf_{T\to\infty} \frac{R_f(T)}{\ln T} \geq \sum_{a:\Delta_a>\epsilon} \frac{f(\Delta_a)}{\mathcal{K}_{\inf}(\nu_a, \mu^* + \epsilon, \mathcal{D})} \ . \tag{3}$$

The proof uses the techniques of (Garivier, Ménard, and Stoltz 2019) and can be found in Appendix B. Specifically, choosing $\epsilon = 0$ leads to the bound for the standard regret (Burnetas and Katehakis 1996). As anticipated, both the lenient regret and the number of samples from arms with large gaps decrease as ϵ increases. This justifies our intuition that removing the penalty from ϵ-optimal arms enables algorithms to reduce the exploration of arms with $\Delta_a > \epsilon$.

The fact that the bounds decrease with ϵ leads to another interesting conclusion – any algorithm that matches the lower bound for some ϵ is *not* consistent for any $\epsilon' < \epsilon$, since it breaks the lower bound for ϵ'. This specifically holds for the standard regret and implies that there is no 'free lunch' – achieving the optimal lenient regret for some $\epsilon > 0$ leads to non-logarithmic standard regret.

Surprisingly, the lower bound is sub-logarithmic when $\mu^* > 1 - \epsilon$. To see this, notice that in this case, there is no distribution $\nu \in \mathcal{D}$ such that $\mathbb{E}[\nu] > \mu^* + \epsilon$, and thus $\mathcal{K}_{\inf}(\nu_a, \mu^* + \epsilon, \mathcal{D}) = \infty$. Intuitively, if the rewards are bounded in $[0,1]$ and some arm has a mean $\mu_a > 1 - \epsilon$, playing it can never incur regret. Identifying that such an arm exists is relatively easy, which leads to low lenient regret. Indeed, we will later present an algorithm that achieves constant regret in this regime.

Finally, and as with most algorithms, we will focus on the set of all problems with rewards bounded in $[0,1]$. In this case, the denominator in Equation (3) is bounded by $\mathcal{K}_{\inf}(\nu_a, \mu^* + \epsilon, \mathcal{D}) \geq d(\mu_a, \mu^* + \epsilon)$ (e.g., by applying Lemma 1 of (Garivier, Ménard, and Stoltz 2019)), and equality holds when the arms are Bernoulli-distributed. Since our results should also hold for Bernoulli arms, our upper bound will similarly depend on $d(\mu_a, \mu^* + \epsilon)$.

Thompson Sampling for Lenient Regret

In this section, we present a modified TS algorithm that can be applied with ϵ-gap functions. W.l.o.g., we assume that the rewards are Bernoulli-distributed, i.e., $X_t \in \{0,1\}$; otherwise, the rewards can be randomly rounded (see (Agrawal and Goyal 2012) for further details). To derive the algorithm, observe that the lower bound of Theorem 1 approaches zero as the optimal arm becomes closer to $1-\epsilon$. Specifically, the lower bound behaves similarly to the regret of the vanilla TS with rewards scaled to $[0, 1-\epsilon]$. On the other hand, if the optimal arm is above $1-\epsilon$, we would like to give it a higher priority, so the regret in this case will be sub-logarithmic. This motivates the following ϵ-TS algorithm, presented in Algorithm 1: denote by $\theta_a(t)$, the sample from the posterior of arm a at round t, and recall that TS algorithm choose arms by $a_t \in \arg\max_a \theta_a(t)$. For any arm with $\hat{\mu}_a(t) \leq 1-\epsilon$, we fix its posterior to be a scaled Beta distribution, such that the range of the posterior is $[0, 1-\epsilon]$, but its mean (approximately) remains $\hat{\mu}_a(t)$ (lines 7-9). If $\hat{\mu}_a(t) > 1-\epsilon$, we set the posterior to $\theta_a(t) = \hat{\mu}_a(t) > 1-\epsilon$ (line 5), which gives this arm a higher priority than any arm with $\hat{\mu}_a(t) \leq 1-\epsilon$. Notice that ϵ-TS *does not* depend on the specific ϵ-gap function. Intuitively, this is since it suffices to match the number of suboptimal plays in Equation (2), that only depends on ϵ. The algorithm enjoys the following asymptotic lenient regret:

Algorithm 1 ϵ-TS for Bernoulli arms

```
 1: Initialize $N_a(1)=0$, $S_a(1)=0$ and $\hat{\mu}_a(1)=0,\ \forall a\in[K]$
 2: for $t = 1, \dots, T$ do
 3:     for $a = 1 \dots, K$ do
 4:         if $\hat{\mu}_a(t) > 1-\epsilon$ then
 5:             $\theta_a(t) = \hat{\mu}_a(t)$
 6:         else
 7:             $\alpha_a(t) = \left\lfloor \frac{S_a(t)}{1-\epsilon} \right\rfloor + 1$
 8:             $\beta_a(t) = N_a(t) + 2 - \alpha_a(t)$
 9:             $\theta_a(t) = (1-\epsilon)Y$ for $Y \sim \mathrm{Beta}(\alpha_a(t), \beta_a(t))$
10:         end if
11:     end for
12:     Play $a_t \in \arg\max_a \theta_a(t)$ and observe the reward $X_t$
13:     Set $N_{a_t}(t+1) = N_{a_t}(t)+1,\ S_{a_t}(t+1) = S_{a_t}(t)+X_t,$
14:     $\hat{\mu}_{a_t}(t+1) = \frac{S_{a_t}(t+1)}{N_{a_t}(t+1)}$ (arms $a \neq a_t$ are unchanged)
15: end for
```

Theorem 2. *Let f be an ϵ-gap function. Then, the lenient regret of ϵ-TS w.r.t. f is*

$$\limsup_{T\to\infty} \frac{R_f(T)}{\ln T} \leq \sum_{a:\Delta_a>\epsilon} \frac{f(\Delta_a)}{d\left(\frac{\mu_a}{1-\epsilon}, \frac{\mu^*}{1-\epsilon}\right)} \tag{4}$$

$$\leq 4(1-\epsilon) \sum_{a:\Delta_a>\epsilon} \frac{f(\Delta_a)}{d(\mu_a, \mu^* + \epsilon)} \ . \tag{5}$$

Moreover, if $\mu^* > 1-\epsilon$, then $R_f(T) = \mathcal{O}(1)$.

The proof can be found in the following section. In our context, the $\mathcal{O}$ notation hides constants that depend on the mean of the arms and ϵ. Notice that Theorem 2 matches the lower bound of Theorem 1 for the set of all bounded distributions (and specifically for Bernoulli arms), up to an absolute constant. Notably, when $\mu^* > 1-\epsilon$, we prove that the regret is constant, and not only sub-logarithmic, as the lower bound suggests. Specifically in this regime, an algorithm can achieve constant lenient regret by identifying an arm with a mean greater than $1-\epsilon$ and exploiting it. However, the algorithm does not know whether such an arm exists, and if there is no such arm, a best arm-identification scheme will perform poorly. Our algorithm naturally identifies such arms when they exist, while maintaining good lenient regret otherwise. Similarly, algorithms such as of (Bubeck, Perchet, and Rigollet 2013) cannot be applied to achieve constant regret, since they require knowing the value of the optimal arm, which is even a stronger requirement than knowing that $\mu^* > 1-\epsilon$.

Comparison to MAB algorithms: Asymptotically optimal MAB algorithms sample suboptimal arms according to the lower bound, i.e., for any suboptimal arm a,

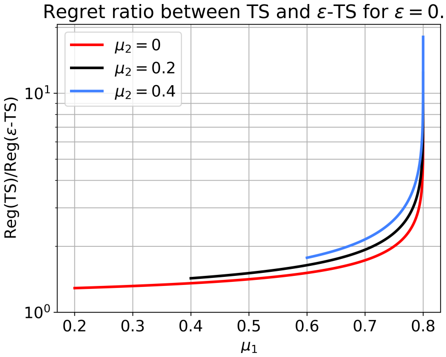

Figure 2: Ratio between the asymptotic lenient regret bounds of TS and ϵ-TS for two-armed problems with $\epsilon = 0.2$, as a function of the optimal arm μ_1.

$\limsup_{T\to\infty} \frac{N_a(T)}{\ln T} \le \frac{1}{d(\mu_a,\mu^*)}$. This, in turn, leads to a lenient regret bound of

$$\limsup_{T\to\infty} \frac{R_f(T)}{\ln T} \le \sum_{a:\Delta_a>\epsilon} \frac{f(\Delta_a)}{d(\mu_a,\mu^*)} \tag{6}$$

that holds for both the vanilla TS (Kaufmann, Korda, and Munos 2012) and KL-UCB (Garivier and Cappé 2011). First notice that the bound of Equation (4), that depends on $d\left(\frac{\mu_a}{1-\epsilon}, \frac{\mu^*}{1-\epsilon}\right)$, strictly improves the bounds for the standard algorithms (see Appendix E.4 for further details). Moreover, ϵ-TS achieves constant regret when $\mu^* > 1-\epsilon$, and its regret quickly diminishes when approaching this regime. This comes in contrast to standard MAB algorithms, that achieve logarithmic regret in these regimes. To illustrate the improvement of ϵ-TS, in comparison to standard algorithms, we present the ratio between the asymptotic bounds of Equations (6) and (4) in Figure 2.

Before presenting the proof, we return to the ϵ-gap function $f(\Delta) = \Delta \cdot \mathbb{1}\{\Delta > \epsilon\}$. Recall that this function leads to the standard regret when all gaps are larger than ϵ. Thus, our algorithm can be applied in this case to greatly improve the performance (from the bound of Equation (6) to the bound of Equation (4)), even in terms of the standard regret.

Regret Analysis

In this section, we prove the regret bound of Theorem 2. For the analysis, we assume w.l.o.g. that the arms are sorted in a decreasing order and all suboptimal arms have gaps $\Delta_a > \epsilon$, i.e. $\mu^* = \mu_1 \ge \mu_1 - \epsilon > \mu_2 \ge \cdots \ge \mu_K$. If there are additional arms with gaps $\Delta_a \le \epsilon$, playing them will cause no regret and the overall lenient regret will only decrease (see Appendix D.1 or Appendix A in (Agrawal and Goyal 2012) for further details). We also assume that $\epsilon < 1$, as otherwise $f(\Delta_a) = 0$ for all $a \in [K]$. Under these assumptions, we now state a more detailed bound for the lenient regret, that also includes a finite-time behavior:

Theorem 3. *Let f be an ϵ-gap function. If $\mu_1 > 1-\epsilon$, there exists some constants $b = b(\mu_1,\mu_2,\epsilon) \in (0,1)$, $C_b = C_b(\mu_1,\mu_2,\epsilon)$ and $L_1 = L_1(\mu_1,\epsilon,b)$ such that*

$$\begin{aligned} R_f(T) &\le \sum_{a=2}^{K} \frac{f(\Delta_a)}{d(1-\epsilon,\mu_a)} \\ &\quad + \max_a f(\Delta_a)\left(C_b + L_1 + \frac{\pi^2/6}{d(1-\epsilon,\mu_1)}\right) \\ &= \mathcal{O}(1)\ . \end{aligned} \tag{7}$$

If $\mu_1 \le 1-\epsilon$, then for any $c > 0$, there exist additional constants $L_2 = L_2(b,\epsilon)$ and $x_{a,c} = x_{a,c}(\mu_1,\mu_a,\epsilon)$ such that for $\eta(t) = \max\left\{\mu_1-\epsilon, \mu_1 - 2\sqrt{\frac{6\ln t}{t^b}}\right\}$,

$$\begin{aligned} R_f(T) &\le (1+c)^2 \sum_{a=2}^{K} f(\Delta_a) \max_{t\in[T]}\left\{\frac{\ln t}{d\left(\frac{\mu_a}{1-\epsilon}, \frac{\eta(t)}{1-\epsilon}\right)}\right\} \\ &\quad + \sum_{a=2}^{K} f(\Delta_a)\left(2 + \frac{1}{c} + \frac{1}{d(x_{a,c},\mu_a)}\right) \\ &\quad + \max_a f(\Delta_a)(C_b + L_2 + 6)\ . \end{aligned} \tag{8}$$

Proof. We decompose the regret similarly to (Kaufmann, Korda, and Munos 2012) and show that with high probability, the optimal arm is sampled polynomially, i.e., $N_1(t) = \Omega(t^b)$ for some $b \in (0,1)$. Formally, let $\eta(t)$ be some function such that $\mu_1 - \epsilon \le \eta(t) < \mu_1$ for all $t \in [T]$, and for brevity, let $f_{\max} = \max_a f(\Delta_a)$. Also, recall that the lenient regret is defined as $R_f(T) = \mathbb{E}\left[\sum_{t=1}^{T} f(\Delta_t)\right]$. Then, the lenient regret can be decomposed to

$$\begin{aligned} R_f(T) &= \sum_{t=1}^{T} \mathbb{E}[f(\Delta_t)(\mathbb{1}\{\theta_1(t) > \eta(t)\} + \mathbb{1}\{\theta_1(t) \le \eta(t)\})] \\ &\le \sum_{t=1}^{T} \mathbb{E}[f(\Delta_t)\mathbb{1}\{\theta_1(t) > \eta(t)\}] + f_{\max} \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}\{\theta_1(t) \le \eta(t)\}] \\ &= \sum_{t=1}^{T}\sum_{a=2}^{K} f(\Delta_a)\mathbb{E}[\mathbb{1}\{a_t = a, \theta_1(t) > \eta(t)\}] \\ &\quad + f_{\max} \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}\{\theta_1(t) \le \eta(t)\}]\ . \end{aligned}$$

Replacing the expectations of indicators with probabilities and dividing the second term to the case where a_1 was sufficiently and insufficiently sampled, we get

$$\begin{aligned} R_f(T) &\le \sum_{a=2}^{K} f(\Delta_a) \underbrace{\sum_{t=1}^{T} \Pr\{a_t = a, \theta_1(t) > \eta(t)\}}_{(A)} \\ &\quad + \underbrace{f_{\max} \sum_{t=1}^{T} \Pr\left\{\theta_1(t) \le \eta(t), N_1(t) > (t-1)^b\right\}}_{(B)} \\ &\quad + \underbrace{f_{\max} \sum_{t=1}^{T} \Pr\left\{N_1(t) \le (t-1)^b\right\}}_{(C)}\ . \end{aligned} \tag{9}$$

The first part of the proof consists of bounding term (C), i.e., showing that the optimal arm is sampled polynomially with high probability. We do so in the following proposition:

Proposition 2. *There exist constants $b = b(\mu_1, \mu_2, \epsilon) \in (0, 1)$ and $C_b = C_b(\mu_1, \mu_2, \epsilon) < \infty$ such that*

$$\sum_{t=1}^{T} \Pr\Big\{N_1(t) \le (t-1)^b\Big\} \le C_b \enspace .$$

The proof follows the lines of Proposition 1 in (Kaufmann, Korda, and Munos 2012) and can be found in Appendix C. To bound (A) and (B), we divide the analysis into two cases: $\mu_1 > 1 - \epsilon$ and $\mu_1 \le 1 - \epsilon$.

First case: $\mu_1 > 1 - \epsilon$.

In this case, we fix $\eta(t) = 1 - \epsilon$. For (A), observe that if $a_t = a$ and $\theta_1(t) > 1 - \epsilon$, then $\theta_a(t) > 1 - \epsilon$, which also implies that $\hat{\mu}_a(t) > 1 - \epsilon$ (to see this, notice that if $\hat{\mu}_a(t) \le 1-\epsilon$, then $\theta_a(t) = (1-\epsilon)Y \le 1-\epsilon$). However, since $\Delta_a > \epsilon$ for all $a \ne 1$, all suboptimal arms have means $\mu_a < 1 - \epsilon$. Thus, when $N_a(t)$ becomes large, the probabilities in (A) quickly diminish and this term can be bounded by constant. Formally, we write

$$\begin{aligned}\sum_{t=1}^{T} &\Pr\{a_t = a, \theta_1(t) > 1 - \epsilon\} \\ &\le \sum_{t=1}^{T} \Pr\{a_t = a, \theta_a(t) > 1 - \epsilon\} \\ &= \sum_{t=1}^{T} \Pr\{a_t = a, \hat{\mu}_a(t) > 1 - \epsilon\}\end{aligned}$$

and bound this term using the following lemma (see Appendix D.2 for the proof):

Lemma 3. *For any arm $a \in [K]$, if $x > \mu_a$, then*

$$\sum_{t=1}^{T} \Pr\{a_t = a, \hat{\mu}_a(t) > x\} \le \frac{1}{d(x, \mu_a)} \enspace .$$

Similarly, in (B), $\theta_1(t) \le 1-\epsilon$ implies that $\hat{\mu}_1(t) \le 1-\epsilon$, and since $N_1(t)$ is large, this event has a low probability. We formalize this intuition in Lemma 4, whose proof can be found in Appendix D.3.

Lemma 4. *Assume that $\mu_1 > 1 - \epsilon$, and for any $b \in (0, 1)$, let $L_1(\mu_1, \epsilon, b)$ such that for all $t \ge L_1(\mu_1, \epsilon, b)$, it holds that $(t-1)^b \ge \frac{2 \ln t}{d(1-\epsilon, \mu_1)} + 1$. Then,*

$$\begin{aligned}\sum_{t=1}^{T} &\Pr\Big\{\theta_1(t) \le 1 - \epsilon, N_1(t) > (t-1)^b\Big\} \\ &\le L_1(\mu_1, \epsilon, b) + \frac{\pi^2/6}{d(1-\epsilon, \mu_1)} \enspace .\end{aligned}$$

Substituting both lemmas and Proposition 2 into Equation (9) leads to Equation (7).

Second case: $\mu_1 \le 1 - \epsilon$.

For this case, we fix $\eta(t) = \max\Big\{\mu_1 - \epsilon, \mu_1 - 2\sqrt{\frac{6 \ln t}{(t-1)^b}}\Big\}$. To bound (A), we adapt the analysis of (Agrawal and Goyal 2013a) and decompose this term into two parts: (i) the event where the empirical mean $\hat{\mu}_a(t)$ is far above μ_a, and (ii) the event where $\hat{\mu}_a(t)$ is close to μ_a and $\theta_a(t)$ is above $\eta(t)$. Doing so leads to Lemma 5, whose proof is in Appendix D.4:

Lemma 5. *Assume that $\mu_1 \le 1 - \epsilon$ and $\eta(t) \in [\mu_1 - \epsilon, \mu_1)$ for all $t \in [T]$. Then, for any $c > 0$,*

$$\begin{aligned}\sum_{t=1}^{T} \Pr\{a_t = a, \theta_1(t) > \eta(t)\} \le (1+c)^2 \max_{t \in [T]} &\left\{ \frac{\ln t}{d\Big(\frac{\mu_a}{1-\epsilon}, \frac{\eta(t)}{1-\epsilon}\Big)} \right\} \\ &+ 2 + \frac{1}{c} + \frac{1}{d(x_{a,c}, \mu_a)} \enspace ,\end{aligned}$$

where $x_{a,c} \in (\mu_a, \mu_1 - \epsilon)$ is such that

$$d\bigg(\frac{x_{a,c}}{1-\epsilon}, \frac{\mu_1 - \epsilon}{1-\epsilon}\bigg) = \frac{1}{1+c} d\bigg(\frac{\mu_a}{1-\epsilon}, \frac{\mu_1 - \epsilon}{1-\epsilon}\bigg).$$

For (B), we provide the following lemma (see Appendix D.5 for the proof):

Lemma 6. *Assume that $\mu_1 \le 1 - \epsilon$ and let $\eta(t) = \max\Big\{\mu_1 - \epsilon, \mu_1 - 2\sqrt{\frac{6 \ln t}{(t-1)^b}}\Big\}$. Also, let $L_2(b, \epsilon) \ge 2$ such that for all $t \ge L_2(b, \epsilon)$, it holds that $\eta(t) > \mu_1 - \epsilon$. Then,*

$$\sum_{t=1}^{T} \Pr\Big\{\theta_1(t) \le \eta(t), N_1(t) > (t-1)^b\Big\} \le L_2(b, \epsilon) + 6$$

Substituting both lemmas and Proposition 2 into (9) results with Equation (8) and concludes the proof of Theorem 3. □

Proof sketch of Theorem 2. It only remains to prove the asymptotic rate of Theorem 2, using the finite-time bound of Theorem 3. To do so, notice that the denominator in Equation (8) asymptotically behaves as $d\Big(\frac{\mu_a}{1-\epsilon}, \frac{\mu_1}{1-\epsilon}\Big)$, which leads to the bound of Equation (4). On the other hand, the denominator of Equation (5) depends on $d(\mu_a, \mu_1 + \epsilon)$. We prove that when $\Delta_a > \epsilon$, these two quantities are closely related:

Lemma 7. *For any $\epsilon \in \big[0, \frac{1}{2}\big)$, any $p \in [0, 1 - 2\epsilon)$ and any $q \in [p + \epsilon, 1 - \epsilon)$,*

$$d\bigg(\frac{p}{1-\epsilon}, \frac{q}{1-\epsilon}\bigg) \ge \frac{1}{4(1-\epsilon)} d(p, q + \epsilon) \enspace .$$

The proof of this lemma can be found in Appendix E.2. This immediately leads to the desired asymptotic rate, but for completeness, we provide the full proof of the theorem in Appendix D.6. □

Experiments

In this section, we present an empirical evaluation of ϵ-TS. Specifically, we compare ϵ-TS to the vanilla TS on two different gap functions: $f(\Delta) = \Delta$, which leads to the standard regret, and the hinge function $f(\Delta) = \max\{\Delta - \epsilon, 0\}$. All evaluations were performed for $\epsilon = 0.2$ over $50,000$ different seeds and are depicted in Figure 3. We also refer the readers to Appendix F, where additional statistics of the simulations are presented, alongside additional tests that were omitted due to space limits. We tested 4 different scenarios – when the optimal arm is smaller or larger than $1 - \epsilon$ (left and right columns, respectively), and when the minimal gap is larger or smaller than ϵ (top and bottom rows, respectively). Importantly, when

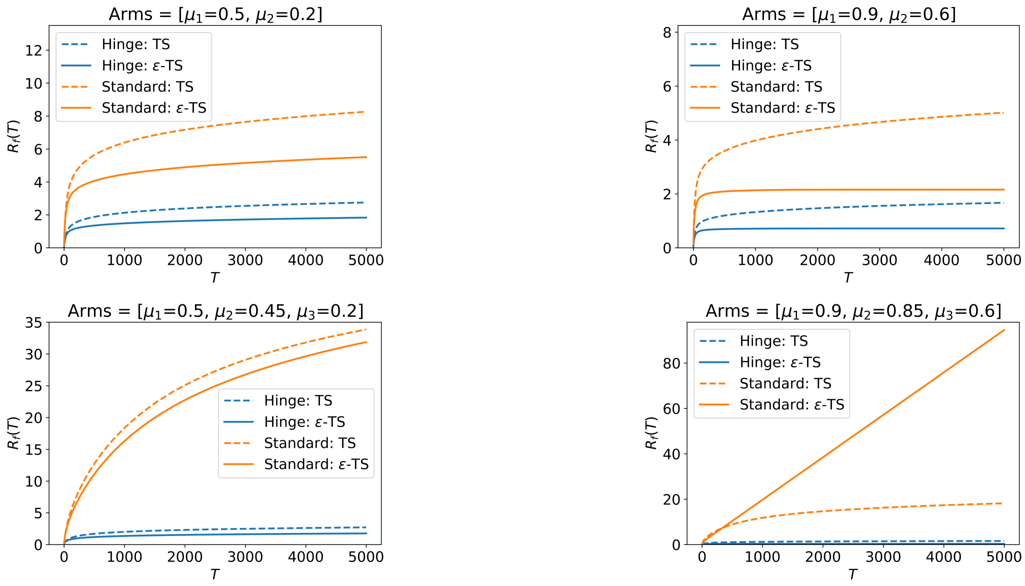

Figure 3: Evaluation of ϵ-TS and vanilla TS with $\epsilon = 0.2$ and Bernoulli rewards. 'Hinge' is the ϵ-gap function $f(\Delta) = \max\{\Delta - \epsilon, 0\}$ and 'Standard' is the 0-gap function $f(\Delta) = \Delta$, which leads to the standard regret. Top row – the minimal gap is $\Delta_2 = 0.3 > \epsilon$; therefore, ϵ-TS enjoys performance guarantees also for the standard regret. Bottom row – the minimal gap is $\Delta_2 = 0.05 < \epsilon$; thus, the standard regret $f(\Delta) = \Delta$ is not an ϵ-gap function, and ϵ-TS has no guarantees for this case.

the minimal gap is larger than ϵ, the standard regret can be written using the ϵ-gap function $f(\Delta) = \Delta \cdot \mathbb{1}\{\Delta > \epsilon\}$. Indeed, one can observe that when $\Delta_a > \epsilon$ for all suboptimal arms, ϵ-TS greatly improves the performance, in comparison to the vanilla TS. Similarly, when $\mu^* > 1 - \epsilon$, the lenient regret of ϵ-TS converges to a constant, as can be expected from Theorem 2. On the other hand, the lenient regret of the vanilla TS continues to increase.

Next, we move to simulations where the suboptimality gap is smaller than ϵ. In such cases, the standard regret cannot be represented as an ϵ-gap function, and ϵ-TS is expected to perform worse on this criterion than the vanilla TS. Quite surprisingly, when $\mu^* = 0.5$, ϵ-TS still surpasses the vanilla TS. In Appendix F, we show that TS beats ϵ-TS only after $20,000$ steps. On the other hand, when $\mu^* = 0.9$, the standard regret of ϵ-TS increases linearly. This is since with finite probability, the algorithm identifies that $\mu_2 = 0.85 > 1 - \epsilon$ at a point where the empirical mean of the optimal arm is smaller than $1 - \epsilon$. Then, the algorithm only exploits $a = 2$ and will never identify that $a = 1$ is the optimal arm. Nonetheless, we emphasize that ϵ-TS still outperforms the vanilla TS in terms of the lenient regret, as can be observed for the hinge-function.

To conclude this section, the simulations clearly demonstrate the tradeoff when optimizing the lenient regret: when near-optimal solutions are adequate, then the performance can be greatly improved. On the other hand, in some cases, it leads to major degradation in the standard regret.

Summary and Future Work

In this work, we introduced the notion of lenient regret w.r.t. ϵ-gap functions. We proved a lower bound for this setting and presented the ϵ-TS algorithm, whose performance matches the lower bound, up to a constant factor. Specifically, we showed that the ϵ-TS greatly improves the performance when a lower bound on the gaps is known. Finally, we performed an empirical evaluation that demonstrates the advantage of our new algorithm when optimizing the lenient regret.

We believe that our work opens up many interesting directions. First, while we suggest a TS algorithm for our settings, it is interesting to devise its UCB counterpart. Moreover, there are alternative ways to define ϵ-gap functions that should be explored, e.g., functions that do not penalize arms with mean larger than $\mu^* \cdot (1 - \epsilon)$ (multiplicative leniency). This can also be done by borrowing other approximation concepts from best arm identification. For example, not penalizing arms that exceed some threshold (as in good arm identification (Kano et al. 2019)), or not penalizing the choice of any one of the top m of the arms (Chaudhuri and Kalyanakrishnan 2017).

We also believe that the concept of lenient regret criteria can be extended to many different settings. It is especially relevant when problems are large, e.g., in combinatorial problems (Chen et al. 2016a), and can also be extended to reinforcement learning (Sutton and Barto 2018). Notably, and as previously stated, there is some similarity between the ϵ-gap function $f(\Delta) = \mathbb{1}\{\Delta > \epsilon\}$ and the sample-complexity criterion in RL (Kakade 2003), and our analysis might allow proving new results for this criterion.

Finally, we explored the notion of lenient regret for stochastic MABs. Another possible direction is adapting the lenient regret to adversarial MABs, and potentially for online learning. In these settings, the convergence rates are typically $\mathcal{O}(\sqrt{T})$, and working with weaker notions of regret might lead to logarithmic convergence rates.

Acknowledgments

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

References

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

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

Agrawal, S.; and Goyal, N. 2013b. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 127–135.

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

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

Bubeck, S.; Perchet, V.; and Rigollet, P. 2013. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, 122–134.

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

Chapelle, O.; and Li, L. 2011. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*, 2249–2257.

Chaudhuri, A. R.; and Kalyanakrishnan, S. 2017. PAC Identification of a Bandit Arm Relative to a Reward Quantile. In *AAAI*, volume 17, 1977–1985.

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

Dann, C.; and Brunskill, E. 2015. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, 2818–2826.

Dann, C.; Lattimore, T.; and Brunskill, E. 2017. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, 5713–5723.

Even-Dar, E.; Mannor, S.; and Mansour, Y. 2002. PAC bounds for multi-armed bandit and Markov decision processes. In *International Conference on Computational Learning Theory*, 255–270.

Gabillon, V.; Ghavamzadeh, M.; and Lazaric, A. 2012. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, 3212–3220.

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

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

Kakade, S. M. 2003. *On the sample complexity of reinforcement learning*. Ph.D. thesis, University College London.

Kano, H.; Honda, J.; Sakamaki, K.; Matsuura, K.; Nakamura, A.; and Sugiyama, M. 2019. Good arm identification via bandit feedback. *Machine Learning* 108(5): 721–745.

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

Korda, N.; Kaufmann, E.; and Munos, R. 2013. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, 1448–1456.

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

Lattimore, T.; Hutter, M.; Sunehag, P.; et al. 2013. The sample-complexity of general reinforcement learning. In *Proceedings of the 30th International Conference on Machine Learning*. Journal of Machine Learning Research.

Lattimore, T.; and Munos, R. 2014. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems*, 550–558.

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

Mannor, S.; and Tsitsiklis, J. N. 2004. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research* 5(Jun): 623–648.

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58(5): 527–535.

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

Slivkins, A.; et al. 2019. Introduction to Multi-Armed Bandits. *Foundations and Trends® in Machine Learning* 12(1-2): 1–286.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

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.

Wang, S.; and Chen, W. 2018. Thompson Sampling for Combinatorial Semi-Bandits. In *International Conference on Machine Learning*, 5114–5122.