Improving Thompson Sampling via Information Relaxation for Budgeted Multi-armed Bandits

Woojin Jeong
wjddnwls5824@kaist.ac.kr
Department of Industrial & Systems Engineering
KAIST, Daejeon, Republic of Korea

Seungki Min
skmin@kaist.ac.kr
Department of Industrial & Systems Engineering
KAIST, Daejeon, Republic of Korea

Abstract

We consider a Bayesian budgeted multi-armed bandit problem, in which each arm consumes a different amount of resources when selected and there is a budget constraint on the total amount of resources that can be used. Budgeted Thompson Sampling (BTS) offers a very effective heuristic to this problem, but its arm-selection rule does not take into account the remaining budget information. We adopt *Information Relaxation Sampling* framework that generalizes Thompson Sampling for classical K-armed bandit problems, and propose a series of algorithms that are randomized like BTS but more carefully optimize their decisions with respect to the budget constraint. In a one-to-one correspondence with these algorithms, a series of performance benchmarks that improve the conventional benchmark are also suggested. Our theoretical analysis and simulation results show that our algorithms (and our benchmarks) make incremental improvements over BTS (respectively, the conventional benchmark) across various settings including a real-world example.

1 Introduction

As an intuitive and efficient heuristic algorithm for sequential decision-making tasks in unknown environments, Thompson Sampling (TS) (Thompson, 1933) has been enjoying a huge success in practice and adopted in recommendation systems (Chapelle & Li, 2011), A/B testing (Graepel et al., 2010), the online advertisement (Graepel et al., 2010; Agarwal, 2013), reinforcement learning (Osband et al., 2013), etc. Built upon online Bayesian inference framework, TS takes an action optimized to model parameters randomly drawn from the posterior distribution at each decision epoch. This simple procedure, called *posterior sampling*, finds a surprisingly proper balance between exploitation and exploration, and is proven to achieve optimality (Agrawal & Goyal, 2012; Russo & Van Roy, 2014).

However, the posterior sampling procedure only considers the current level of model uncertainty, not considering the future consequences of individual actions. This often critically affects the performance of TS, particularly when the value of exploration needs to be taken into account carefully – for example, when there are an excessive number of arms (Russo & Van Roy, 2022) when the arms have different noise variances (Kirschner & Krause, 2018; Min et al., 2020), or when the exploration is restricted due to a budget constraint, the situation formulated as a *budgeted multi-armed bandit* (MAB) (Ding et al., 2013; Xia et al., 2015).

In the budgeted MAB, playing an arm yields a random reward and incurs a deterministic/random cost at the same time, and no more play can be made once the playe runs out of budget. This setting has been introduced to model online bidding optimization in sponsored search (Amin et al., 2012; Tran-Thanh et al., 2014), and on-spot instance bidding in cloud computing (Agmon Ben-Yehuda et al., 2013). The algorithms such as KUBE (Tran-Thanh et al., 2012), UCB-BV1/BV2 (Ding et al., 2013), PD-BwK (Badanidiyuru et al., 2013), i/c/m-UCB, b-Greedy (Xia et al., 2017), and BTS

(Xia et al., 2015) have been proposed and analyzed. Budgeted Thompson Sampling (BTS), as an immediate extension of TS for budgeted MAB, is considered as a baseline algorithm to be fixed in this work. Although it significantly outperforms the other algorithms, it still does not consider the remaining budget information when making a decision, and hence suffers from the aforementioned issue.

To overcome this shortcoming, we adopt the *Information Relaxation Sampling* (IRS) framework, recently suggested by Min et al. (2019) for classical Bayesian K-armed bandit problems. Generalizing the concept of posterior sampling, the IRS framework suggests a class of algorithms which optimize their actions to a randomly generated future scenario (not just model parameters) in a careful consideration of the belief dynamics of Bayesian learners.

Our contributions are threefold: First, by applying the IRS framework to the budgeted MAB setting, we develope a series of algorithms that can exploit the specific details of the problem instance such as budget information. Without introducing any auxiliary parameter, they easily achieve the state-of-the-art performance. In our numerical experiment, the improvement over BTS can be as large as 75% in terms of reduction in regret.

Second, we obtain as byproducts a series of upper bounds on the maximal performance that can be achieved in the given problem instance. This series of upper bounds also improve the conventional one commonly used in the definition of Bayesian regret, and turn out to be useful to see how much additional improvement can be made.

Finally, we extend IRS to random cost settings by making two levels of extensions. As a relatively simpler extension, we allow IRS policies to sample the mean cost values from their posterior distributions and then solve inner problems as if these sampled values are the ground truth, i.e., the idea of IRS is applied only to rewards but not to costs. As a more complicated extension, we can make IRS policies to sample all future cost realizations and then solve more complex inner problems that additionally consider how much the decision maker will learn about the cost distributions, i.e., the idea of IRS is applied to both rewards and costs. Our numerical experiment shows that these extensions of IRS policies indeed offer sequential improvements over BTS as expected. And it show that the more complicated extension outperforms the simple extension.

Throughout this paper, we will focus on explaining two specific algorithms, namely, IRS.FH and IRS.V-Zero, instead of describing the general framework.

2 Problem Formulation and Preliminaries

We consider a Bayesian budgeted MAB problem with K arms and a resource budget B. A problem instance can be specified by a tuple $\big(K, B, (c_a, \mathcal{R}_a, \Theta_a, \mathcal{P}_a, \mathcal{Y}_a, y_{a,0})_{a\in[K]}\big)$ which will be described in a greater detail below.

Rewards and costs. Let $\mathcal{A} = [K]$ be the set of arms, among which the decision maker (DM) can play one in each time period. The stochastic reward that the DM earns from the n^{th} pull of arm a is represented with a nonnegative random variable $R_{a,n}$, and we assume that its distribution is given by $\mathcal{R}_a(\theta_a)$:

$$R_{a,n} \sim \mathcal{R}_a(\theta_a), \quad \forall n = 1, 2, \dots,$$

where $\theta_a \in \Theta_a$ is the *unknown parameter* that the DM aims to learn. Given θ_a, the rewards $R_{a,1}, R_{a,2}, \dots$ are independent.

Whenever arm a is played, it also incurs a deterministic cost,[1] denoted by $c_a \in \mathbb{N}$ (i.e., consumes c_a units of resources deterministically). The total amount of resources that the DM can use is limited by $B \in \mathbb{N}$, and the DM's goal is to maximize the expected total reward within this budget constraint.

[1] In development and analyses of our suggested algorithms, we primarily focus on the deterministic cost setting. The main ideas naturally extend to random cost setting. See § 4.

Bayesian framework. In the Bayesian framework, the unknown parameter θ_a is treated as a random variable and we assume that its prior distribution is given by $\mathcal{P}_a(y_{a,0})$, i.e.,

$$\theta_a \sim \mathcal{P}_a(y_{a,0}),$$

where the hyperparameter $y_{a,0} \in \mathcal{Y}_a$, which we call (initial) *belief*, specifies the prior distribution.

As a Bayesian learner, the DM's belief about θ_a will be updated according to the Bayes' rule whenever a new reward realization from the arm a is observed. To describe the belief dynamics explicitly, we introduce a *Bayesian update function* $\mathcal{U}_a : \mathcal{Y}_a \times \mathbb{R}^+ \to \mathcal{Y}_a$. That is, after playing the arm a for the first time, the belief is updated from $y_{a,0}$ to $y_{a,1} \triangleq \mathcal{U}_a(y_{a,0}, R_{a,1})$ and then the posterior distribution of θ_a can be written as $\mathcal{P}_a(y_{a,1})$. We accordingly define $y_{a,n}$ be the belief that the DM will have after playing the arm n times, i.e., $y_{a,n} \triangleq \mathcal{U}_a(y_{a,n-1}, R_{a,n})$ for $n = 1, 2, \ldots$.

Mean reward estimates. We denote the unknown mean reward of arm a by $\mu_a(\theta_a)$ as a real-valued function of parameter θ_a:

$$\mu_a(\theta_a) \triangleq \mathbb{E}[R_{a,n}|\theta_a].$$

Let us denote its n-sample (Bayesian) estimate by $\hat{\mu}_{a,n}(\cdot)$ as a real-valued function of first n reward realizations: abbreviating $(R_{a,1}, \ldots, R_{a,n})$ as $R_{a,1:n}$, we define

$$\hat{\mu}_{a,n}(R_{a,1:n}\ ; y_{a,0}) \triangleq \mathbb{E}_{\theta_a \sim \mathcal{P}_a(y_{a,0})}[\mu_a(\theta_a)|R_{a,1:n}],$$

which represents the expected performance of arm a inferred from its first n reward realizations, or equivalently, the predictive mean reward of arm a that the DM would believe after playing the arm n times.

These mean-reward metrics μ_a and $\hat{\mu}_{a,n}$ will be repeatedly used throughout the paper. The reason why we define μ_a and $\hat{\mu}_{a,n}$ as functions is to clarify their dependencies on the random variables and to utilize their functional form when developing algorithms later. To help understanding, we make the following remark.

Remark 1 *By Strong Law of Large Numbers, we have*

$$\lim_{n\to\infty} \hat{\mu}_{a,n}(R_{a,1:n}\ ; y_{a,0}) = \mu_a(\theta_a), \quad a.s.,$$

which says that, in terms of mean-reward estimation, knowing the parameter is equivalent to having an infinite number of observations. Also, for any n and k, it holds that

$$\hat{\mu}_{a,n+k}(R_{a,1:n+k}\ ; y_{a,0}) = \hat{\mu}_{a,k}(R_{a,n+1:n+k}\ ; y_{a,n}),$$

which says that making an inference using $n+k$ samples given an initial belief is equivalent to making an inference using the later k samples after updating the belief using the former n samples.

Policy and performance. Let π be the DM's policy, and A_t be the arm played by π at time t. The reward that the DM earns at time t can be written as

$$r_t \triangleq R_{A_t, n_{A_t,t}} \quad \text{where} \quad n_{a,t} \triangleq \sum_{s=1}^{t} \mathbf{1}\{A_s = a\}.$$

Here, $n_{a,t}$ counts the number of times that arm a has been played up to time t. An admissible policy π should decide A_t based only on the information revealed prior to time t, $(A_s, r_s)_{s=1}^{t-1}$.

Besides, playing the arm A_t consumes c_{A_t} units of resources. To describe the budget constraint explicitly, we introduce a stopping time τ representing the first time that the cumulative cost exceeds the given budget, i.e.,

$$\tau \triangleq \min\left\{ t : \sum_{s=1}^{t} c_{A_s} > B \right\}.$$

Only the rewards realized before time τ are counted, so the total reward collected by the DM can be written as $\sum_{t=1}^{\tau-1} r_t$. As a trivial upper bound on τ, we introduce $T_{\max} \triangleq \max_{a\in\mathcal{A}} \{\lfloor B/c_a \rfloor + 1\}$.

We denote by $V(\pi)$ the expected performance of policy π in a given MAB instance:

$$V(\pi) \triangleq \mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau-1} r_t\right] = \mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau-1} \mu_{A_t}(\theta_{A_t})\right].$$

Here, the expectation operator takes into account the randomness of the policy (if randomized like BTS), the reward realizations $R_{1:K,1:T_{\max}}$, and the parameter realizations $\theta_{1:K}$. Note that $V(\pi)$ can be alternatively represented as $\mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau-1} \mu_{A_t}(\theta_{A_t})\right]$ by the law of total expectation, since $\mathbb{E}[r_t|A_t, \theta_{1:K}] = \mu_{A_t}(\theta_{A_t})$.

Performance bound and regret. A quantity W is said to be a *performance bound* if $W \geq V(\pi)$ for any policy π.

As a performance bound commonly used in the MAB literature, W^{BTS} is defined as[2]

$$W^{\mathrm{BTS}} \triangleq \mathbb{E}\left[B \times \max_{a\in\mathcal{A}} \frac{\mu_a(\theta_a)}{c_a}\right]. \tag{1}$$

This quantity represents the expected performance of the clairvoyant fractional solution: when the player knows the parameters $\theta_{1:K}$ in advance, it is optimal for him to play the arm $a^\star$ with the largest reward-to-cost ratio $\mu_a(\theta_a)/c_a$, (fractionally) $B/c_{a^\star}$ times in a row, which will yield the total reward of $\mathbb{E}\left[\mu_{a^\star}(\theta_{a^\star}) \times B/c_{a^\star}\right]$ $(= W^{\mathrm{BTS}})$ in average. Clearly, no policy can perform better than this clairvoyant player, and therefore, W^{BTS} is an upper bound on the maximal achievable performance for the given MAB instance.

A performance bound W is said to be *tighter* than the other W' if $W \leq W'$. A tighter bound provides a more precise quantification of the hardness of a particular MAB instance, and can better serve as a performance benchmark.

On the other hand, we will later utilize the *Bayesian regret* to visualize and compare the performance of policies, which is defined as

$$\textsc{Regret}(\pi) \triangleq W^{\mathrm{BTS}} - V(\pi).$$

The regret quantifies the suboptimality of a policy, and is non-negative since W^{BTS} is a performance bound. Once we have a performance bound W tighter than W^{BTS}, the gap $W^{\mathrm{BTS}} - W$ will provide a lower bound on the minimal achievable regret (i.e., $\textsc{Regret}(\pi) \geq W^{\mathrm{BTS}} - W$ for any π).

Bayesian optimal policy. In the Bayesian setting, there exists a policy that achieves the maximal performance $V^\star$:

$$V^\star \triangleq \sup_{\pi} V(\pi).$$

Such a *Bayes-optimal policy* and its performance $V^\star$, in theory, can be obtained by solving the Bellman equation (corresponding to an MDP with a state space $\mathcal{Y}_1 \times \ldots \times \mathcal{Y}_K$ and an action space $\mathcal{A}$. See Appendix A for the detail), but they are intractable in most cases.

As motivated in the introduction, our primary goal is to improve the BTS policy in terms of performance, where the Bayes-optimal policy will be our ideal target. Another goal is to improve the performance bound W^{BTS} in terms of tightness, where $V^\star$ will be our ideal target.

3 Algorithms

In this section, we propose a series of policies that improve Budgeted Thompson Sampling (BTS) toward the Bayes-optimal policy by leveraging the idea of *information relaxation sampling*. In parallel,

[2]The naming W^{BTS} is not common in the literature. The motivation for this choice is explained in §3.1.

we argue that there is a performance bound embedded in each of these policies, and accordingly, the performance bounds paired with our proposed policies also improve the performance bound paired with BTS, which is the conventional benchmark W^{BTS}.

3.1 Budgeted Thompson Sampling

As an immediate extension of Thompson Sampling to the budgeted MAB setting, BTS (Xia et al., 2015) utilizes the posterior sampling of the parameters. As described in Algorithm 1, the policy π^{BTS} at each time t draws a random sample of the parameters from the posterior distribution (i.e., $\tilde{\theta}_a^{(t)} \sim \mathcal{P}_a(y_{a,n_{a,t-1}})$ in line 4), and plays the arm with the largest reward-to-cost ratio given the sampled parameters (i.e., $\arg\max_a \mu_a(\tilde{\theta}_a^{(t)})/c_a$ in line 6). After observing the result of the play, it updates the belief about the arm according to the Bayes' rule (line 11), and repeats this procedure until the budget is exhausted.

Algorithm 1 BTS

Input: $K, B, (c_a, \mathcal{R}_a, \Theta_a, \mathcal{P}_a, \mathcal{Y}_a, y_{a,0})_{a\in[K]}$
Procedure:

1: Initialize $t \leftarrow 1$, $B_1 \leftarrow B$, $n_{a,0} \leftarrow 0$ for each $a \in \mathcal{A}$
2: **while** $B_t > 0$ **do**
3: **for** each arm $a \in \mathcal{A}$ **do**
4: Sample $\tilde{\theta}_a^{(t)} \sim \mathcal{P}_a(y_{a,n_{a,t-1}})$
5: **end for**
6: $A_t \leftarrow \arg\max_{a\in\mathcal{A}}\{\mu_a(\tilde{\theta}_a^{(t)})/c_a\}$
7: **if** $B_t < c_{A_t}$ **then**
8: break
9: **else**
10: Play A_t, receive r_t, pay c_{A_t} $(B_{t+1} \leftarrow B_t - c_{A_t})$
11: Update $y_{a,n_{a,t-1}+1} \leftarrow \mathcal{U}_a(y_{a,n_{a,t-1}}, r_t)$, and $n_{a,t} \leftarrow \begin{cases} n_{a,t-1}+1 & \text{for } a = A_t \\ n_{a,t-1} & \text{for } a \neq A_t \end{cases}$
12: **end if**
13: $t \leftarrow t + 1$.
14: **end while**

One can immediately relate this arm-selection rule with the performance bound W^{BTS}, defined in (1). As motivated earlier, the arm $a^\star = \arg\max_{a\in\mathcal{A}}\{\mu_a(\theta_a)/c_a\}$ is the optimal one to play if the parameters are known and the fractional solution is allowed. The policy π^{BTS} mimics such a clairvoyant player's decision by replacing the unknown components $\mu_a(\theta_a)$'s with their randomly generated counterparts $\mu_a(\tilde{\theta}_a^{(t)})$'s. Note that the randomness in this sampling procedure enforces π^{BTS} to deviate from the myopic decision, resulting in explorations.

Although BTS is simple and computationally efficient ($O(K)$ computations per decision), its arm-selection rule does not incorporate the remaining budget information. As an extreme example, if the remaining budget is so small that each arm can be play at most once, it is Bayes-optimal to make the myopic decision, i.e., $A_t \leftarrow \arg\max_{a\in\mathcal{A}}\{\mathbb{E}_{\theta_a\sim\mathcal{P}_a(y_{a,n_{a,t-1}})}[\mu_a(\theta_a)]/c_a\}$. For this reason, BTS often performs unnecessary explorations, particularly near the end of horizon, which motivates next algorithm IRS.FH.

3.2 IRS.FH

Our first proposed algorithm IRS.FH[3] is very similar to BTS but additionally incorporates how many times each arm can be played in the future within the remaining budget. While the belief

[3]IRS stands for Information Relaxation Sampling, and FH stands for Finite Horizon.

updating procedure remains unchanged, IRS.FH implements a slightly different arm-selection rule, which is described in Algorithm 2 (lines 3–7).

Algorithm 2 IRS.FH

Input: $K, B, (c_a, \mathcal{R}_a, \Theta_a, \mathcal{P}_a, \mathcal{Y}_a, y_{a,0})_{a\in[K]}$
Procedure:

1: Initialize $t \leftarrow 1$, $B_1 \leftarrow B$, $n_{a,0} \leftarrow 0$ for each $a \in \mathcal{A}$
2: **while** $B_t > 0$ **do**
3: **for** each arm $a \in \mathcal{A}$ **do**
4: Sample $\tilde{\theta}_a^{(t)} \sim \mathcal{P}_a(y_{a,n_{a,t-1}})$ and $\tilde{R}_{a,i}^{(t)} \sim \mathcal{R}_a(\tilde{\theta}_a^{(t)})$ for $i = 1, \ldots, \lfloor B_t/c_a \rfloor$
5: $\tilde{\hat{\mu}}_{a,\lfloor B_t/c_a \rfloor}^{(t)} \leftarrow \hat{\mu}_{a,\lfloor B_t/c_a \rfloor}(\tilde{R}_{a,1:\lfloor B_t/c_a \rfloor}^{(t)}; y_{a,n_{a,t-1}})$
6: **end for**
7: $A_t \leftarrow \arg\max_{a\in\mathcal{A}} \{\tilde{\hat{\mu}}_{a,\lfloor B_t/c_a \rfloor}^{(t)}/c_a\}$
8: Play A_t and update variables (Algorithm 1 lines 7–13)
9: **end while**

Policy $\pi^{\text{IRS.FH}}$. More specifically, the policy $\pi^{\text{IRS.FH}}$ at each time samples not only the parameters $\tilde{\theta}_a^{(t)}$'s but also all future rewards $\tilde{R}_{a,i}^{(t)}$'s (line 4). Here, $\tilde{R}_{a,i}^{(t)}$ represents the sampled reward realization associated with the future i^{th} play of arm a, where $i \le \lfloor B_t/c_a \rfloor - 1$ since the arm a can be updated at most $\lfloor B_t/c_a \rfloor - 1$ times when the remaining budget is B_t. Given these sampled future rewards, it computes the future $(\lfloor B_t/c_a \rfloor - 1)$-sample mean-reward estimate $\tilde{\hat{\mu}}_{a,\lfloor B_t/c_a \rfloor - 1}^{(t)}$, i.e., the belief that we would have if we allocate all remaining budget to the arm a and the sampled future rewards indeed realize. Finally, the arm with the largest reward-to-cost $\tilde{\hat{\mu}}_{a,\lfloor B_t/c_a \rfloor - 1}^{(t)}/c_a$ is selected: this is almost identical to the arm-selection rule of BTS except that $\tilde{\hat{\mu}}_{a,\lfloor B/c_a \rfloor - 1}^{(t)}$ is used instead of $\mu_a(\tilde{\theta}_a^{(t)})$.

In other words, $\pi^{\text{IRS.FH}}$ finds the best arm given a finite-number of randomly synthesized future observations. By simulating the future belief changes using the sampled future rewards, it naturally takes into account how much we can learn in the future: when a smaller amount of budget is remaining, fewer future rewards will be sampled, and thus the future belief will less deviate from the current belief, which makes $\pi^{\text{IRS.FH}}$ more myopic, desirably.

Let us examine the Beta-Bernoulli case for example: when the current belief is $y_a = (\alpha_a, \beta_a)$ and the remaining budget is B, $\tilde{\hat{\mu}}_{a,\lfloor B/c_a \rfloor - 1}$ can be expressed as

$$\tilde{\hat{\mu}}_{a,\lfloor B/c_a \rfloor} = \frac{\alpha_a + \sum_{i=1}^{\lfloor B/c_a \rfloor - 1} \tilde{R}_{a,i}}{\alpha_a + \beta_a + \lfloor B/c_a \rfloor - 1}.$$

Note that, when B is small, $\tilde{\hat{\mu}}_{a,\lfloor B/c_a \rfloor - 1} \approx \frac{\alpha_a}{\alpha_a+\beta_a} = \mathbb{E}_{\theta_a \sim \text{Beta}(\alpha_a,\beta_a)}[\mu_a(\theta_a)]$ which leads to the myopic decision (i.e., exploitation), and when B is large, $\tilde{\hat{\mu}}_{a,\lfloor B/c_a \rfloor - 1} \approx \frac{1}{\lfloor B/c_a \rfloor - 1} \sum_{i=1}^{\lfloor B/c_a \rfloor - 1} \tilde{R}_{a,i} \approx \mu_a(\tilde{\theta}_a^t)$ which leads to the BTS's decision. Like this, the degree of exploration is naturally adjusted depending on the amount of remaining budget, mitigating the over-exploration issue that BTS suffers from.

We also remark that IRS.FH can be computationally efficient as much as BTS. Observe that in the above example $\sum_{i=1}^{\lfloor B/c_a \rfloor - 1} \tilde{R}_{a,i}$ is distributed with Binomial$(\lfloor B/c_a \rfloor - 1, \tilde{\theta}_a)$, and therefore $\tilde{\hat{\mu}}_{a,\lfloor B/c_a \rfloor - 1}$ can be computed via a single random number generation without sampling $\tilde{R}_{a,i}$'s one by one. Such a trick is applicable to more general situations where the reward distribution belongs to natural exponential family, and both IRS.FH and BTS requires $O(K)$ computations per decision.

Bound $W^{\text{IRS.FH}}$. We can motivate a new performance bound $W^{\text{IRS.FH}}$ that is associated with $\pi^{\text{IRS.FH}}$. Analogously to the way that we relate π^{BTS} with W^{BTS}, we define

$$W^{\text{IRS.FH}} \triangleq \mathbb{E}\left[B \times \max_{a\in\mathcal{A}} \frac{\hat{\mu}_{a,\lfloor B/c_a \rfloor - 1}(R_{a,1:\lfloor B/c_a \rfloor - 1}\,; y_{a,0})}{c_a}\right].$$

Compared to W^{BTS}, this bound implicitly postulates another type of clairvoyant player who knows $\hat{\mu}_{a,\lfloor B/c_a \rfloor}$ instead of $\mu_a(\theta_a)$. In the task of identifying the best arm, the finite-sample mean-reward estimate $\hat{\mu}_{a,\lfloor B/c_a \rfloor}$ is less informative than the true mean-reward $\mu_a(\theta_a)$ (recall Remark 1). Therefore, the clairvoyant player informed with $\hat{\mu}_{a,\lfloor B/c_a \rfloor - 1}$ cannot perform better than the one informed with $\mu_a(\theta_a)$, which implies that $W^{\mathrm{IRS.FH}}$ is tighter than W^{TS}. We show in Theorem 1 that $W^{\mathrm{IRS.FH}}$ is indeed a valid performance bound and improves W^{BTS} in terms of tightness (i.e., $W^{\mathrm{BTS}} \geq W^{\mathrm{IRS.FH}} \geq V^\star$).

On the other hand, the value of $W^{\mathrm{IRS.FH}}$ can be computed via sample averaging scheme, i.e., by repeatedly computing the term inside the expectation with respect to randomly generated $\hat{\mu}_{a,\lfloor B/c_a \rfloor} - 1$'s. This procedure can be simply implemented by reusing the code of $\pi^{\mathrm{IRS.FH}}$ (lines 3–7).

3.3 IRS.V-Zero

We consequently propose our next algorithm, IRS.V-Zero[4], that further improves IRS.FH by solving a more complicated optimization problem in each time period. It takes into account not only how many times each arm can be played, but also how the belief changes over the course of future plays.

Algorithm 3 IRS.V-Zero

Input: $K, B, (c_a, \mathcal{R}_a, \Theta_a, \mathcal{P}_a, \mathcal{Y}_a, y_{a,0})_{a \in [K]}$

Procedure:

1: Initialize $t \leftarrow 1$, $B_1 \leftarrow B$, $n_{a,0} \leftarrow 0$ for each $a \in \mathcal{A}$
2:
3: **while** $B_t > 0$ **do**
4: **for** each arm $a \in \mathcal{A}$ **do**
5: Sample $\tilde{\theta}_a^{(t)} \sim \mathcal{P}_a(y_{a,n_{a,t-1}})$ and $\tilde{R}_{a,i}^{(t)} \sim \mathcal{R}_a(\tilde{\theta}_a^{(t)})$ for $i = 1, \ldots, \lfloor B_t/c_a \rfloor$
6: **for** $i = 1, \ldots, \lfloor B_t/c_a \rfloor$ **do**
7: $\tilde{\hat{\mu}}_{a,i}^{(t)} \leftarrow \hat{\mu}_{a,i}(\tilde{R}_{a,1:i}^{(t)}; y_{a,n_{a,t-1}})$
8: **end for**
9: **end for**
10:

$$\text{Solve } \tilde{n}_{1:K}^\star \leftarrow \operatorname*{arg\,max}_{\tilde{n}_{1:K} \in \mathcal{N}(B_t)} \sum_{a=1}^{K} \sum_{i=1}^{\tilde{n}_a} \tilde{\hat{\mu}}_{a,i-1}^{(t)}, \text{ where } \mathcal{N}(B_t) \triangleq \{(\tilde{n}_1, \ldots, \tilde{n}_K); \sum_{a=1}^{K} c_a \tilde{n}_a \leq B_t\}$$

11: $A_t \leftarrow \arg\max_{a \in \mathcal{A}} \tilde{n}_a^\star$
12: Play A_t and update variables (Algorithm 1 lines 7–13)
13: **end while**

Policy $\pi^{\mathbf{IRS.V\text{-}Zero}}$. The pseudo-code is given in Algorithm 3. The policy $\pi^{\mathrm{IRS.V\text{-}Zero}}$ samples the entire future reward realizations just like $\pi^{\mathrm{IRS.FH}}$ does, and computes all future finite-sample estimates $\tilde{\hat{\mu}}_{a,i}^{(t)}$ for $i = 1, 2, \ldots, \lfloor B_t/c_a \rfloor$ sequentially. And then it solves a knapsack-like optimization problem (line 10) so as to determine how many times each arm should be played in the sampled future: with the nonnegative decision variables $\tilde{n}_1, \ldots, \tilde{n}_K$, it solves

$$\text{maximize} \sum_{a=1}^{K} \sum_{i=1}^{\tilde{n}_a} \tilde{\hat{\mu}}_{a,i-1}^{(t)} \text{ subject to } \sum_{a=1}^{K} \tilde{n}_a c_a \leq B_t. \tag{2}$$

Given the optimal solution $(\tilde{n}_1^\star, \ldots, \tilde{n}_K^\star) \in \mathbb{N}^K$, it actually plays the arm with the largest $\tilde{n}_a^\star$ (line 11) with an arbitrary tie-breaking rule.

The optimization problem (2) is to find the "optimal allocation of the remaining budget across the arms". In its objective, the term $\tilde{\hat{\mu}}_{a,i-1}$ represents the predictive mean reward of the future i^{th} play

[4] V-zero stands for the penalty associated with setting the prior of the information relaxation penalty discussed in §3.4 to 0.

(predicted with the future belief right after the $(i-1)^{\text{th}}$ play), and the term $\sum_{i=1}^{\tilde{n}_a} \tilde{\hat{\mu}}_{a,i-1}$ represents the expected cumulative reward that can be obtained from the next $\tilde{n}_a$ plays of arm a.

Compared to the optimization problem that IRS.FH solves ($B \times \max_a \{\tilde{\hat{\mu}}_{a,\lfloor B/c_a \rfloor}/c_a\}$), it additionally takes into account how the belief will change over the course of the future plays, not just what the final belief will be. This also reflects the fact that the player has to allocate i plays in order to obtain the estimate $\hat{\mu}_{a,i}$. By considering more carefully this future belief dynamics, $\pi^{\text{IRS.V-Zero}}$ achieves a better balance between exploitation and exploration than $\pi^{\text{IRS.FH}}$ does. However, the optimization problem (2) requires $O(KBT_{\max})$ computations to solve, which is considerably slower than IRS.FH.

Bound $W^{\text{IRS.V-Zero}}$. Focusing on the optimization problem (2), we immediately obtain the following performance bound:

$$W^{\text{IRS.V-Zero}} \triangleq \mathbb{E}\left[\max_{n_{1:K} \in \mathcal{N}_B} \sum_{a=1}^{K} \sum_{i=1}^{n_a} \hat{\mu}_{a,i-1}\right],$$

where $\mathcal{N}_B \triangleq \{(n_1, \ldots, n_K); \sum_{a=1}^{K} c_a n_a \leq B\}$, and $\hat{\mu}_{a,i-1}$ hides its dependency on $R_{a,1:i-1}$ and $y_{a,0}$ for better presentation. In Theorem 1, we show that $W^{\text{IRS.V-Zero}}$ further improves $W^{\text{IRS.FH}}$.

3.4 Generalization

Note that all of three policies, π^{BTS}, $\pi^{\text{IRS.FH}}$, and $\pi^{\text{IRS.V-Zero}}$, share the following structure in common: they in each time period (i) randomly generate future information via posterior sampling, (ii) optimize their decision to this randomly generated future via solving a deterministic optimization problem (referred to as *inner problem*), (iii) play an arm according to the optimized decision, and update the belief according to Bayes' rule. Their corresponding performance bounds, W^{BTS}, $W^{\text{IRS.FH}}$, and $W^{\text{IRS.V-Zero}}$, can be obtained by solving the same inner problems, not with the sampled future realizations, but with the true future realizations.

The *information relaxation sampling* (IRS) framework formally generalizes this structure with the notion of information relaxation penalties. Deferring its detailed description to Appendix A, we briefly remark that IRS unifies BTS and the Bayesian optimal policy (OPT) into a single framework, and also includes IRS.FH, IRS.V-Zero, and IRS.V-EMax as special cases that interpolate between BTS and OPT.

Each policy-bound pair is characterized by inner optimization problem: from BTS to OPT, they introduce increasingly complicated optimization problems, becoming more considerate but more computationally costly. We indeed observe and (partly) prove that these policies achieve increasingly better performance and these performance bounds achieve increasingly better tightness.

In addition, we also implement and evaluate IRS.INDEX policy, which, strictly speaking, does not belong to IRS framework (it does not have a corresponding performance bound). It internally utilizes IRS.V-EMax to obtain a random approximation of the Gittins index. See Appendix A for the detail.

4 Extension to Random Cost

We have so far developed our framework for deterministic cost setting. In this section, we extend IRS framework to random cost setting, in which each arm consumes a random amount of resource whenever played and this random cost is drawn from an unknown distribution that we also aim to learn. More specifically, the stochastic cost that the DM pays for the n^{th} pull of arm a is represented with a nonnegative random variable $C_{a,n}$. Every notation is analogously defined for costs, while we use superscript ^{c} (or ^{r}) to represent the parameters/variables related to costs (or rewards, respectively): e.g., the distribution of $C_{a,n}$ is given by $\mathcal{C}_a(\theta_a^c)$, where θ_a^c is the unknown parameter for which we have a prior $\mathcal{P}_a^c(y_{a,0}^c)$.

IRS algorithms can be extended to random cost in multiples ways. We here explore two ideas — a simple extension that uses the sampled mean cost, and a bit more complicated extension that uses the sampled future cost realizations and introduces additional penalties.

Simple extension As described in Xia et al. (2015), BTS applied to the random cost setting draws the parameters θ_a^c's from the posterior, and selects the arm with the largest mean-reward-to-mean-cost ratio: i.e., $\arg\max_a \mu_a^r(\tilde{\theta}_a^r)/\mu_a^c(\tilde{\theta}_a^c)$. Analogously, we motivate simple extensions of IRS policies that solve the same inner problems to the deterministic cost setting but use $\mu_a^c(\tilde{\theta}_a^c)$ instead of c_a.

Extension with additional penalties In the deterministic cost setting, we have motivated IRS polices by relaxing the information constraint imposed on reward realizations. Similarly, we can consider to relax the information constraint imposed on cost realizations. That is, we can let a policy to sample the future cost realizations in addition to the future reward realizations and solve some deterministic optimization problem with respect to this sampled future but in the presence of penalties for letting the DM exploit the future information. A penalty function suitable for IRS.V-Zero can be designed as follows.[5]

The penalty function of IRS.V-Zero is given by

$$z_t^{\text{IRS.V-Zero}}(a_{1:t}, \omega) \triangleq r_t(a_{1:t}, \omega) - \mathbb{E}_y[r_t(a_{1:t}, \omega) | H_{t-1}(a_{1:t-1}, \omega)].$$

This penalizes the DM for knowing the future reward realizations, and similarly, we can add an extra term that penalizes the DM for knowing the future cost realizations:

$$\begin{aligned} z_t^{\text{IRS.V-Zero}}(a_{1:t}, \omega) \triangleq r_t(a_{1:t}, \omega) &- \mathbb{E}_{y^r}[r_t(a_{1:t}, \omega) | H_{t-1}(a_{1:t-1}, \omega)] \\ &- \lambda\Big(c_t(a_{1:t}, \omega) - \mathbb{E}_{y^c}[c_t(a_{1:t}, \omega) | H_{t-1}(a_{1:t-1}, \omega)]\Big). \end{aligned}$$

Here, $\lambda \in \mathbb{R}$ supposedly captures the additional benefit that the DM can earn by knowing the actual cost realization at time t instead of its expected value. A natural choice of λ will be the dual variable associated with the budget constraint of the inner problem that IRS.V-Zero solves, i.e., $\lambda = \max_a \mu(\theta_a^r)/\mu(\theta_a^c)$, the quantity reflects the additional benefit that the DM can earn when one unit of resource is additionally given. We consider an extended version of IRS.V-Zero policy that uses its sampled value, i.e., $\tilde{\lambda} = \max_a \mu(\tilde{\theta}_a^r)/\mu(\tilde{\theta}_a^c)$, resulting in the following inner problem:

$$\max_{n_1,\ldots,n_K} \sum_{a=1}^{K} \sum_{i=1}^{n_a} \left\{\hat{\mu}_{a,i-1}^r + \tilde{\lambda}(\tilde{C}_{a,i} - \hat{\mu}_{a,i-1}^c)\right\} \text{ s.t. } \sum_{a=1}^{K} \sum_{i=1}^{n_a} \tilde{C}_{a,i} \leq B.$$

5 Analysis

We first provide a theoretical result showing that the performance bounds W^{BTS} and $W^{\text{IRS.FH}}$ proposed in §3 are valid upper bounds on the maximal achievable performance and incrementally tighter than the conventional benchmark.

Theorem 1 (Monotonicity of performance bounds) *For any Bayesian budgeted MAB, we have*

$$W^{BTS} \geq W^{IRS.FH} \geq W^{IRS.V\text{-}Zero} \geq V^\star.$$

The formal proof of Theorem 1 is given in Appendix C. We briefly sketch the main idea as follows. Recall that each of these bounds represents the maximal performance that can be achieved by a clairvoyant player who has an access to some additional information that is supposed to be unknown, and therefore, it should be greater than $V^\star$, the maximal performance of the non-clairvoyant player.

[5] We extended the penalty functions not only for IRS.V-Zero but also to IRS.V-EMax and IRS.INDEX policy. The detailed procedure of two extensions is implemented in Appendix B.

In this line of thought, the gap $W - V^\star$ can be understood as a quantity that measures how much additional benefit can be extracted by exploiting the additional information, which should decrease when less useful information is additionally given. This explains the monotonicity $W^{\mathrm{BTS}} \geq W^{\mathrm{IRS.FH}} \geq W^{\mathrm{IRS.V\text{-}Zero}}$, which is formally proven via Jensen's inequality.

On the other hand, the improvements in the performance bounds ($W^{\mathrm{BTS}} \to W^{\mathrm{IRS.FH}} \to W^{\mathrm{IRS.V\text{-}Zero}}$) naturally imply the improvements in their corresponding policies ($\pi^{\mathrm{BTS}} \to \pi^{\mathrm{IRS.FH}} \to \pi^{\mathrm{IRS.V\text{-}Zero}}$). Recall that each of these policies mimics the behavior of the clairvoyant player using the self-generated future information, i.e., it plays an arm that would have been selected by the one who optimistically believes that the sampled future information is the ground truth. The gap $W - V^\star$ now can be translated as a quantity that measures how overly optimistic the corresponding policy will behave. Hence, the policy associated with a tighter performance bound is less likely to make a decision that is overly optimized to a particular realization of future information, and avoids over-explorations more effectively.

We indeed observe in all our numerical experiments that the suggested policies monotonically improve BTS in terms of performance, i.e., $V(\pi^{\mathrm{BTS}}) \leq V(\pi^{\mathrm{IRS.FH}}) \leq V(\pi^{\mathrm{IRS.V\text{-}Zero}})$. However, proving this monotonicity is very challenging, so we instead investigate the gaps between the performance of these policies and their corresponding performance bounds, and establish upper bounds on these gaps.

Theorem 2 (Suboptimality gap) *Consider a Bayesian budgeted MAB such that $\mathcal{R}_a$ is a natural exponential family distribution specified by a log-partition function $A_a(\theta_a)$ and $\mathcal{P}_a$ is given by its conjugate prior whose density function is of the form $\exp(\xi_a\theta_a - \nu_a A_a(\theta))$. Suppose that all the log-partition functions are L-smooth, i.e., $\frac{d^2}{d\theta_a^2}A_a(\theta_a) \leq L$, $\forall \theta_a \in \Theta_a$, and $\nu_a = \nu$, $\forall a \in \mathcal{A}$. Then, for any $B \geq 2\max\{c_1, \ldots, c_K\}$, we have*

$$
\begin{aligned}
W^{BTS} - V(\pi^{BTS}) &\leq 2\sqrt{L}\left[\frac{1}{\sqrt{\nu}} + \sqrt{2\log T_{max}}\left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT_{max}}\right)\right], \\
W^{IRS.FH} - V(\pi^{IRS.FH}) &\leq 2\sqrt{L}\left[\frac{1}{\sqrt{\nu}} + \sqrt{2\log T_{max}}\left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT_{max}} - \frac{1}{3}\sqrt{\frac{T_{max}}{K}}\right)\right], \\
W^{IRS.V\text{-}Zero} - V(\pi^{IRS.V\text{-}Zero}) &\leq \sqrt{L}\left[\frac{1}{\sqrt{\nu}} + \sqrt{2\log T_{max}}\left(\frac{K}{\sqrt{\nu}} + 2\sqrt{KT_{max}} - \frac{1}{3}\sqrt{\frac{T_{max}}{K}}\right)\right],
\end{aligned}
$$

where $T_{max} \triangleq \max_{a\in\mathcal{A}}\{\lfloor B/c_a \rfloor + 1\}$.

Theorem 2 considers Bayesian budgeted MABs with natural exponential family distributions, which include the Beta-Bernoulli case ($L = 1/2$, $\nu = \alpha+\beta$) and the Beta-Binomial case ($L = m/2$, $\nu = (\alpha+\beta)/m$). While all these suboptimality gaps have the same asymptotic order of $O(\sqrt{KT_{\max}\log T_{\max}})$, this result shows that IRS.FH and IRS.V-Zero make incremental improvements over BTS in the additional term and in the leading coefficient. The proof is given in Appendix D.

Note that our analysis aligns closely with the regret lower bound analysis and the algorithm's regret upper bound analysis typically conducted in the MAB literature. Theorem 1 provides a tighter lower bound $V^\star$ compared to the lower bound W^{BTS} presented in other budgeted MAB-related studies. Theorem 2 highlights improvements in the suboptimality gap, distinct from the regret upper bound $W^{\mathrm{BTS}} - V(\pi)$ noted in other budgeted MAB literature. The observed reduction in the suboptimality gap may be due to enhancements in $V(\pi)$, although it remains somewhat ambiguous whether these improvements are predominantly due to W^π or $V(\pi)$. This ambiguity makes direct comparisons of $V(\pi)$ values challenging and renders the result less robust. Nevertheless, experimental evidence substantiates that notable improvements are also achieved in $V(\pi)$.

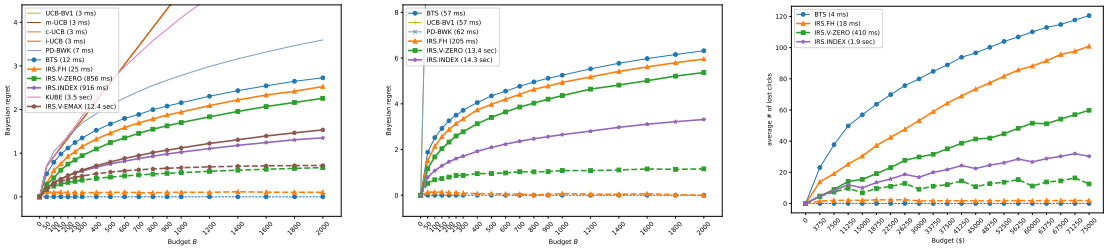

Figure 1: From left to right, simulation results in (a) the Beta-Bernoulli MAB with two arms, (b) the Beta-Bernoulli MAB with five arms, and (c) the Beta-Binomial MAB as a real-world example arising in online advertisement business.

6 Numerical Experiments

We demonstrate the effectiveness of our proposed policies and performance bounds through numerical simulations. We consider deterministic cost setting with three MAB instances[6] – (a) the Beta-Bernuolli MAB with two arms, (b) the Beta-Bernuolli MAB with five arms, and (c) the Beta-Binomial MAB with six arms as a real-world example arising in the online advertisement business. In each setting, we evaluate the empirical performance of IRS policies as well as their corresponding performance bounds, and also provide a comparison with KUBE (Tran-Thanh et al., 2012), UCB-BV1 (Ding et al., 2013), i/c/m-UCB (Xia et al., 2017), and a modified version of PD-BwK (Badanidiyuru et al., 2013) as competing benchmarks.

Figure 1 visualizes the simulation results in these three settings where the x-axes represent the budget B. The solid-line curves report the regret of the policies ($W^{\mathrm{BTS}} - V(\pi)$), and the dashed-line curves report the regret lower bounds obtained with the performance bounds ($W^{\mathrm{BTS}} - W$). The run time of each policy is reported in the legend, representing the average time to complete a single run of simulation.

Beta-Bernoulli MABs. We first examine a Beta-Bernoulli MAB instance with $K = 2$, $(c_1, c_2) = (10, 20)$, and $\alpha_a = \beta_a = 1, \forall a \in \mathcal{A}$, and report the result of 50,000 runs of simulation in Figure 1(a). When $B = 2,000$, BTS outperform all competing benchmarks by a large margin, from 32% (BTS's regret vs. PD-BwK's regret) up to 228% (BTS's regret vs. i/c/m-UCB & UCB-BV1's regret). Our proposed policies even further improve BTS: IRS.FH, IRS.V-Zero, IRS.V-EMax, and IRS.INDEX policies, respectively, achieve 8%, 18%, 44%, and 51% improvement over BTS in terms of reduction in regret. Furthermore, we can infer from the regret lower bound $W^{\mathrm{BTS}} - W^{\mathrm{IRS.V\text{-}EMax}}$ (brown dashed-line curve) that no policy can achieve an improvement more than 74%, highlighting that IRS.INDEX policy is near optimal.

We next examine a Beta-Bernoulli MAB instance with $K = 5$, $c_{1:5} = (2, 3, 10, 19, 20)$, and $\alpha_a = \beta_a = 1, \forall a \in \mathcal{A}$, and report the result of 20,000 runs of simulation in Figure 1(b). IRS.V-EMax is excluded due to its computational inefficiency. The gaps between BTS and other benchmarks are even larger, and IRS.FH, IRS.V-Zero, and IRS.INDEX policies, respectively, achieve 6%, 15%, and 48% improvement over BTS.

Application to online advertisement budget allocation. We examine a Beta-Binomial MAB instance that represents a bandit task encountered by a company who wants to optimally allocate his marketing budget across a number of ad campaigns with unknown click-through-rates (CTRs). More specifically, the arms represent the campaigns available to this company, and playing an arm a means that the company decides to spend c_a dollars on the campaign a on the next day which will create

[6] We also show that IRS algorithms are sufficiently scalable for random cost setting through numerical simulation. See Appendix E for the detail.

m_a impressions. The company's goal is to maximize the total number of clicks using a marketing budget B dollars where the prior distribution of CTR can be approximated by $\text{Beta}(\alpha_a, \beta_a)$.

The problem constants $(c_a, m_a, \alpha_a, \beta_a)$'s were chosen based on iPinYou dataset (Liao et al., 2014), a publicly available real-world dataset containing logs of ad auctions, bids, impressions, clicks, and final conversions. Imagining a region-based marketing strategy, we have empirically estimated average cost-per-impression, average number of daily impressions, and CTRs in six different regions separately, and obtained the values $(c_a, m_a, \alpha_a, \beta_a) = (\$3750, 30204, 12, 14153), (\$7200, 55965, 22, 22950), (\$15000, 120485, 25, 28968), (\$12750, 105148, 34, 44244), (\$2700, 22952, 17, 20977), (\$3300, 29847, 20, 22559)$ for $a = 1, \ldots, 6$, respectively. We simulate the algorithms while varying the budget B from \$3,750 to \$75,000.

As shown in Figure 1(c), IRS.FH, IRS.V-Zero, and IRS.INDEX policies, respectively, achieve 16%, 50%, and 75% improvement over BTS, when the budget is \$75,000. Converted into dollars, the improvement made by IRS.INDEX is valuable as much as \$10,000 approximately. Given that it is impossible to reduce BTS's regret more than 89% (as implied by $W^{\text{BTS}} - W^{\text{IRS.V-Zero}}$), we can conclude that IRS.INDEX is near-optimal.

7 Conclusion

We have proposed a series of algorithms for budgeted MAB that improve Thompson sampling utilizing the information relaxation. In their arm-selection procedure, they simulate Bayesian learner's belief dynamics with respect to the sampled future realizations, and by doing so they can take into account how much the decision maker can learn within the remaining budget constraint. As a byproduct, our framework produces performance bounds that provide better quantifications of possible improvement. While the main ideas are mostly adopted from Min et al. (2019), this paper highlights that the information relaxation technique is particularly effective for budgeted bandit tasks, in which finding an optimal balance between exploration and exploitation is critical.

Our contribution may seem obvious, but it is far from trivial. Existing literature on Budgeted MAB did not consider the use of the remaining budget information at all, and its extension in the context of the IRS framework presented its application in more realistic and appropriate settings. Unlike classic MAB problems, the termination time (the total number of pulls, denoted by stopping time τ in our proof) depends on the sequence of actions, which introduces additional challenges requiring careful theoretical analysis and complicates algorithm implementation.

We further extend the framework to the random cost setting. The adoption of the IRS framework naturally necessitates the inclusion of cost sampling. However, a challenge arises regarding the imposition of an information relaxation penalty on cost in this context. address this challenge, we propose introducing a dual variable for the budget constraint, algorithmically simplifying it to the posterior mean reward-cost ratio. This dual variable concept holds promise for extending the imposition of additional penalties beyond budget constraints, potentially encompassing scenarios such as bandit problems with multiple constraints.

Acknowledgments

This research was supported by the Asian Office of Aerospace R&D (AOARD, Award No.: FA2386-23-1-4122).

References

Deepak Agarwal. Computational advertising: the linkedin way. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 1585–1586, 2013.

Orna Agmon Ben-Yehuda, Muli Ben-Yehuda, Assaf Schuster, and Dan Tsafrir. Deconstructing amazon ec2 spot instance pricing. *ACM Transactions on Economics and Computation (TEAC)*, 1(3):1–20, 2013.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012.

Kareem Amin, Michael Kearns, Peter Key, and Anton Schwaighofer. Budget optimization for sponsored search: Censored learning in mdps. *arXiv preprint arXiv:1210.4847*, 2012.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 207–216. IEEE, 2013.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

Thore Graepel, Joaquin Quinonero Candela, Thomas Borchert, and Ralf Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. Omnipress, 2010.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pp. 592–600. PMLR, 2012.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pp. 358–384. PMLR, 2018.

Hairen Liao, Lingxiao Peng, Zhenchuan Liu, and Xuehua Shen. ipinyou global rtb bidding algorithm competition dataset. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, pp. 1–6, 2014.

Seungki Min, Costis Maglaras, and Ciamac C Moallemi. Thompson sampling with information relaxation penalties. *Advances in Neural Information Processing Systems*, 32, 2019.

Seungki Min, Ciamac C Moallemi, and Daniel J Russo. Policy gradient optimization of thompson sampling policies. *arXiv preprint arXiv:2006.16507*, 2020.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

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

Daniel Russo and Benjamin Van Roy. Satisficing in time-sensitive bandit learning. *Mathematics of Operations Research*, 2022.

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

Long Tran-Thanh, Archie Chapman, Alex Rogers, and Nicholas Jennings. Knapsack based optimal policies for budget–limited multi–armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pp. 1134–1140, 2012.

Long Tran-Thanh, Lampros Stavrogiannis, Victor Naroditskiy, Valentin Robu, Nicholas R Jennings, and Peter Key. Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions. In *uai2014, 30th Conf. on Uncertainty in AI*, 2014.

Yingce Xia, Haifang Li, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Thompson sampling for budgeted multi-armed bandits. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Yingce Xia, Tao Qin, Wenkui Ding, Haifang Li, Xudong Zhang, Nenghai Yu, and Tie-Yan Liu. Finite budget analysis of multi-armed bandit problems. *Neurocomputing*, 258:13–29, 2017.