Bayesian bandits: balancing the exploration-exploitation tradeoff via double sampling

Iñigo Urteaga and Chris H. Wiggins
{inigo.urteaga, chris.wiggins}@columbia.edu

Department of Applied Physics and Applied Mathematics
Data Science Institute
Columbia University
New York City, NY 10027

August 8, 2018

Abstract

Reinforcement learning studies how to balance exploration and exploitation in real-world systems, optimizing interactions with the world while simultaneously learning how the world operates. One general class of algorithms for such learning is the multi-armed bandit setting. Randomized probability matching, based upon the Thompson sampling approach introduced in the 1930s, has recently been shown to perform well and to enjoy provable optimality properties. It permits generative, interpretable modeling in a Bayesian setting, where prior knowledge is incorporated, and the computed posteriors naturally capture the full state of knowledge. In this work, we harness the information contained in the Bayesian posterior and estimate its sufficient statistics via sampling. In several application domains, for example in health and medicine, each interaction with the world can be expensive and invasive, whereas drawing samples from the model is relatively inexpensive. Exploiting this viewpoint, we develop a double sampling technique driven by the uncertainty in the learning process: it favors exploitation when certain about the properties of each arm, exploring otherwise. The proposed algorithm does not make any distributional assumption and it is applicable to complex reward distributions, as long as Bayesian posterior updates are computable. Utilizing the estimated posterior sufficient statistics, double sampling autonomously balances the exploration-exploitation tradeoff to make better informed decisions. We empirically show its reduced cumulative regret when compared to state-of-the-art alternatives in representative bandit settings.

1 Introduction

In a plethora of problems in science and engineering, one needs to decide which action to take next, based on partial information about the options available: a doctor must prescribe a medicine to a patient, a manager must allocate resources to competing projects, an ad serving algorithm must decide where to place ads, etc. In practice, the underlying properties of each choice are only partially known at the time of the decision, but one hopes that the understanding of the caveats involved will improve as time passes.

This set of problems has an illustrative gambling analogy, where a person facing a row of slot machines needs to devise its playing strategy (policy): which arms to play and in which order. The aim is to maximize the expected reward after a certain set of actions. Statisticians have studied this abstraction under the name of the multi-armed bandit problem for decades, e.g., in the seminal works by Robbins (17, 18). The multi-armed bandit setting consists of sequential interactions with the world with rewards that are independent and identically distributed, or the related contextual bandit case, in which the reward distribution depends on different information or 'context' presented with each interaction. It has played an important role in many fields across science and engineering.

Several algorithms have been proposed to overcome the exploration-exploitation tradeoff in such problems, mostly based on heuristics, on upper confidence bounds, or on the Gittins index. From the former, the ϵ-greedy approach (randomly pick an arm with probability ϵ, otherwise be greedy) has become very popular due to its simplicity, while nonetheless retaining often good performance (4). In the latter case, Gittins (10) formulated a method based on computing the optimal strategy for certain types of bandits, where geometrically discounted future rewards are considered. There are several difficulties inherent to the exact computation of the Gittins index and thus, approximations have been developed as well (8). These and other intrinsic challenges of the method have limited its applicability (23).

Lai and Robbins (14) introduced another class of algorithms, based on upper confidence intervals of the expected reward of each arm, for which strong theoretical guarantees were proved (13). Nevertheless, these algorithms might be far from optimal in the presence of dependent and more general reward distributions (21). Bayesian counterparts of UCB-type algorithms have been proposed in (11), where they show it provides an unifying framework for other variants of the UCB algorithm for distinctive bandit problems.

Recently, the problem has re-emerged both from a practical (importance in e-commerce and web applications, e.g., (15)) and a theoretical (research on probability matching algorithms and their regret bounds, e.g., (1) and (16)) point of view.

Contributing to this revival was the observation that one of the oldest heuristics to address the exploration-exploitation tradeoff, i.e., Thompson sampling (24, 25), has been empirically proven to perform satisfactorily (see (9) and (22) for details). Contemporaneously, theoretical study established several performance bounds, both for problems with and without context (2, 3, 12, 19, 20).

In this work, we are interested in the randomized probability matching approach, as it connects to the Bayesian learning paradigm. It readily facilitates not only generative and interpretable modeling, but sequential and batch processing algorithm development too.

Specifically, we investigate the benefits of fully harnessing the posteriors obtained via the Bayesian sequential learning process. We hereby avoid distributional assumptions to allow for complicated relationships among action rewards, as long as Bayesian posterior updates are computable.

We explore the benefits of sampling the model posterior to estimate the sufficient statistics that drive randomized probability matching algorithms. Our motivation is cases where sampling from the model posterior is inexpensive relative to interacting with the world, which may be expensive or invasive or, as in the medical application domain, both. The goal is that, with informative posterior sufficient statistics, better decisions can be made, leading to a lower cumulative regret.

We propose a double sampling technique for the multi-armed bandit problem, based on (1) Monte Carlo sampling, to approximate otherwise unsolvable integrals, and (2), a sampling-based arm-selection policy.

The policy is driven by the uncertainty in the learning process, as it favors exploitation when certain about the properties of the arms, exploring otherwise. Due to this autonomous exploration-exploitation balancing technique, the proposed algorithm achieves improved average performance, with important regret reductions.

We formally introduce the problem in Section 2, before providing all the details of our proposed double sampling method in Section 3. The performance of double sampling is compared to the Thompson sampling and Bayes-UCB alternatives in Section 4, and we conclude with final remarks in Section 5.

2 Problem formulation

We mathematically formulate the multi-armed bandit problem as follows. Let $a \in \{1, \cdots, A\}$ indicate the arms of the bandit (possible actions to take), and $f_a(y|\theta)$ the stochastic reward distribution of each arm. For every time instant, the observed reward y_t is independently drawn from the reward distribution corresponding to the played arm. We denote as a_t the arm played at time instant t; $a_{1:t} \equiv (a_1, \cdots, a_t)$ refers to the sequence of arms played up to time t, and similarly, $y_{1:t} \equiv (y_1, \cdots, y_t)$ to the sequence of observed rewards.

In the multi-armed bandit setting one must decide, based on observed rewards $y_{1:t}$ and actions $a_{1:t}$, which arm to play next in order to maximize rewards. Due to the stochastic nature of the rewards, their expectation under the arm's distribution is the statistic of interest. We denote each arm's expected reward as $\mu_a(\theta) = \mathbb{E}_a\{y|\theta\}$, which is parameterized by the arm-dependent parameters θ.

When the properties of the arms (i.e., their parameters) are known, one can readily determine the optimal selection policy, i.e.,

$$a^*(\theta) = \underset{a}{\operatorname{argmax}}\, \mu_a(\theta) \; . \tag{1}$$

However, the optimal solution for the multi-armed bandit is only computable in closed form in very few special cases (5, 10), and it fails to generalize to more realistic reward distributions and scenarios (21). The biggest challenge occurs when the parameters are unknown, as one might end up playing the wrong arm forever if incomplete learning occurs (7).

Amongst the different algorithms to overcome these issues, the randomized probability matching, i.e., playing each arm in proportion to its probability of being optimal, is a particularly appealing one. It has shown to be easy to implement, efficient and broadly applicable.

Given the parameters θ, the expected reward of each arm is deterministic and, thus, one must pick the arm with the maximum expected reward

$$\Pr\left[a = a^*_{t+1} | a_{1:t}, y_{1:t}, \theta\right] = \Pr\left[a = a^*_{t+1} | \theta\right] = I_a(\theta), \tag{2}$$

where we use the indicator function

$$I_a(\theta) = \begin{cases} 1, \; \mu_a(\theta) = \max\{\mu_1(\theta), \cdots, \mu_A(\theta)\} \; , \\ 0, \; \text{otherwise} \; . \end{cases} \tag{3}$$

Under random probability matching, the aim is to compute the probability of a given arm a being optimal for the next time instant, $p_{a,t+1} \in [0, 1]$, even with unknown parameters.

Mathematically,

$$\begin{aligned} p_{a,t+1} &\equiv \Pr\left[a = a^*_{t+1} \middle| a_{1:t}, y_{1:t}\right] \\ &\equiv \Pr\left[\mu_a(\theta) = \max\{\mu_1(\theta), \cdots, \mu_A(\theta)\} \middle| a_{1:t}, y_{1:t}\right]. \end{aligned} \tag{4}$$

Note that there is an inherent uncertainty about the unknown properties of the arms, as Eqn. (4) is parameterized by θ. In order to compute a solution to this problem, recasting it as a Bayesian learning problem, where θ is a random variable, is of great help. It allows for computation of posterior and marginal distributions, with direct connection to sampling techniques.

3 Proposed method: double sampling

The multi-armed bandit problem consists of two separate but intertwined tasks: (1) learning about the properties of the arms, and (2) deciding what arm to play next. The problem is sequential in nature, as one makes a decision on which arm to play and learns from the observed reward, one observation at a time.

We cast the multi-armed bandit problem as a sequential Bayesian learning task. By doing so, we capture the full state of knowledge about the world at every time instant. We incorporate any available prior information to the learning process, and update our knowledge about the unknown parameter θ, as we sequentially play arms and observe rewards. This learning can be done both sequentially or in batches, as Bayesian posterior updates are computable for both cases (6).

However, the solution to the probability matching equation in (4) is analytically intractable, so we approximate it via Monte Carlo sampling. For balancing the exploration-exploitation tradeoff, we propose a sampling-based probability matching technique too. The proposed arm-selection policy is a function of the uncertainty in the learning process. The intuition is that we exploit only when certain about the properties of the arms, while we keep exploring otherwise.

We elaborate on the foundations of the proposed double sampling method in the following sections, before presenting it formally in Algorithm 1.

3.1 Bayesian multi-armed bandits

We are interested in computing, after playing arms $a_{1:t}$ and observing rewards $y_{1:t}$, the probability $p_{a,t+1}$ of each arm a being optimal for the next time instant. In practice, one needs to account for the lack of knowledge of each arm's properties, i.e., the unknown parameter θ in Eqn. (4).

We do so by following the Bayesian methodology, where the parameters are considered to be another set of random variables. The uncertainty over the parameters can be accounted for by marginalizing over their probability distribution.

Specifically, we marginalize over the posterior of the parameters after observing rewards and actions up to time t,

$$\begin{aligned} p_{a,t+1} \equiv \Pr\left[a = a^*_{t+1} \middle| a_{1:t}, y_{1:t}\right] &= f(a = a^*_{t+1} | a_{1:t}, y_{1:t}) \\ &= \int f(a = a^*_{t+1} | a_{1:t}, y_{1:t}, \theta) f(\theta | a_{1:t}, y_{1:t}) \mathrm{d}\theta \,. \end{aligned} \tag{5}$$

Given a prior for the parameters $f(\theta)$ and the per-arm reward distribution $f_a(y|\theta)$, one can compute the posterior of each arm's parameters by

$$\begin{aligned} f(\theta|a_{1:t}, y_{1:t}) &\propto f_{a_t}(y_t|\theta) f(\theta|a_{1:t-1}, y_{1:t-1}) \\ &\propto \left[\prod_{\tau=1}^{t} f_{a_\tau}(y_\tau|\theta)\right] f(\theta) \ . \end{aligned} \tag{6}$$

This posterior provides information (with uncertainty) about the characteristics of the arm. Note that the updates can usually be written in both sequential and batch forms. This flexibility is of great help in many practical scenarios, as one can learn from historic observations, as well as process data as it comes.

Even if analytical expressions for the parameter posteriors are available for many models of interest, computing the probability of any given arm being optimal is analytically intractable, due to the nonlinearities induced by the indicator function as in Eqn. (3)

$$p_{a,t+1} = \int f(a = a^*_{t+1}|a_{1:t}, y_{1:t}, \theta) f(\theta|a_{1:t}, y_{1:t}) \mathrm{d}\theta = \int I_a(\theta) f(\theta|a_{1:t}, y_{1:t}) \mathrm{d}\theta \ . \tag{7}$$

3.2 Monte-Carlo integration

We harness the power of Monte Carlo sampling to compute the otherwise analytically intractable integral in Eqn. (7). We obtain a Monte Carlo based random measure approximation to compute estimates of $p_{a,t+1} \in [0, 1]$ as follows:

1. Draw M parameter samples from the updated posterior distribution

$$\theta^{(m)} \sim f(\theta|a_{1:t}, y_{1:t}), \ \ m = \{1, \cdots, M\} \ . \tag{8}$$

2. For each parameter sample $\theta^{(m)}$, compute the expected reward and determine the best arm

$$a^*_{t+1}(\theta^{(m)}) = \underset{a}{\mathrm{argmax}}\, \mu_a(\theta^{(m)}) \ . \tag{9}$$

3. Define the random measure approximation

$$f(a = a^*_{t+1}|a_{1:t}, y_{1:t}) \approx f_M(a = a^*_{t+1}|a_{1:t}, y_{1:t}) \approx \frac{1}{M} \sum_{m=1}^{M} \delta\left(a - a^*_{t+1}(\theta^{(m)})\right), \tag{10}$$

where $\delta(\cdot)$ denotes the Dirac delta function.

4. Estimate the first- and second-order sufficient statistics of $f_M(a = a^*_{t+1}|a_{1:t}, y_{1:t})$, i.e.,

$$\begin{cases} \hat{p}_{a,t+1} = \mathbb{E}\left\{f_M(a = a^*_{t+1}|a_{1:t}, y_{1:t})\right\} = \frac{1}{M}\sum_{m=1}^{M} I_a\left(\theta^{(m)}\right) \ , \\ \hat{\sigma}^2_{a,t+1} = \mathbb{V}\mathrm{ar}\{f_M(a = a^*_{t+1}|a_{1:t}, y_{1:t})\} = \frac{1}{M}\sum_{m=1}^{M}\left(I_a\left(\theta^{(m)}\right) - \hat{p}_{a,t+1}\right)^2 \ . \end{cases} \tag{11}$$

5. Estimate which is the optimal arm and with what probability

$$\begin{cases} \hat{a}^*_{t+1} = \mathrm{argmax}_a\, \hat{p}_{a,t+1} \ , \\ \hat{p}^*_{a,t+1} = \max_a \hat{p}_{a,t+1} \ . \end{cases} \tag{12}$$

3.3 Sampling-based policy

In any bandit setting, given the available information at time t, one needs to decide which arm to play next. A randomized probability matching technique would pick the next arm a with probability $p_{a,t+1}$. On the contrary, a greedy approach would choose the arm with the highest probability of being optimal, i.e., $p^*_{a,t+1}$.

We present an alternative sampling-based probability matching arm-selection policy that finds a balance between these two cases. We rely on the Monte Carlo approximation to Eqn. (7), and leverage the estimated sufficient statistics in Eqn. (11) to balance the exploration-exploitation tradeoff. We draw candidate arm samples from the random measure in Eqn. (10), and automatically adjust the probability matching technique according to the accuracy of this approximation.

The number of candidate arm samples drawn is instrumental for our sampling-based policy. We automatically adjust its value according to the uncertainty on the optimality of each arm, i.e., $\sigma^2_{a,t+1}$. By doing so, we account for the uncertainty of the learning process in the arm-selection policy, dynamically balancing exploration and exploitation.

The number of candidate arm samples to draw is inversely proportional to the probability of not picking the optimal arm. We denote this probability as p_{FA}, which is computed for each arm as

$$p^{(a)}_{FA} = Pr\left(p_{a,t+1} > p^*_{a,t+1}\right) = 1 - F_{p_{a,t+1}}(p^*_{a,t+1}) \,, \tag{13}$$

where $p^*_{a,t+1} = \max_a p_{a,t+1}$. The true cumulative density function $F_{p_{a,t+1}}(\cdot)$ is analytically intractable as well, but we approximate it (based on the central limit theorem guarantees of the MC estimates) with a Gaussian truncated to the CDF's range

$$F_{p_{a,t+1}}(p^*_{a,t+1}) \approx \Phi_{[0,1]}\left(\frac{p^*_{a,t+1} - p_{a,t+1}}{\sigma_{a,t+1}}\right) \,. \tag{14}$$

Since we can not exactly evaluate $p_{a,t+1}$ and $\sigma_{a,t+1}$, we resort to our Monte Carlo estimates in Eqn. (11) instead.

All in all, the proposed sampling policy proceeds as follows:

1. Determine N_{t+1}, the number of candidate arm samples to draw

$$\begin{aligned} N_{t+1} \propto \log\left(\frac{1}{p_{FA}}\right), \;\; p_{FA} = \frac{1}{K-1}\sum_{a \neq \hat{a}^*_{t+1}} p^{(a)}_{FA} \,, \\ p^{(a)}_{FA} \approx 1 - \Phi_{[0,1]}\left(\frac{\hat{p}^*_{a,t+1} - \hat{p}_{a,t+1}}{\hat{\sigma}_{a,t+1}}\right) \,. \end{aligned} \tag{15}$$

2. Draw N_{t+1} candidate arm samples

$$\hat{a}^{(n)}_{t+1} \sim \text{Cat}\left(\hat{p}_{a,t+1}\right), \quad n = 1, \cdots, N_{t+1} \,. \tag{16}$$

3. Pick the most probable optimal arm, given drawn candidate arm samples $\hat{a}^{(n)}_{t+1}$

$$a_{t+1} = \text{Mode}\left(\hat{a}^{(n)}_{t+1}\right), \quad n = 1, \cdots, N_{t+1} \,. \tag{17}$$

By allowing for N_{t+1} to be adjusted based upon the uncertainty of the learning process, we balance the exploration-exploitation tradeoff. We present full details of the proposed double sampling technique in Algorithm 1.

Algorithm 1 Double sampling algorithm

Require: Number of arms A, number of MC samples M, and horizon T

Require: Prior over model parameters $f(\theta)$ and per-arm reward distributions $f_a(y|x,\theta)$

$D = \emptyset$

for $t = 1, \cdots, T$ **do**

Draw M posterior parameter samples

$$\theta_{t+1}^{(m)} \sim f(\theta|a_{1:t}, x_{1:t}, y_{1:t}), \quad m = \{1, \cdots, M\} \tag{18}$$

If applicable, receive context x_{t+1}

for $a = 1, \cdots, A$ **do**

Compute expected reward,
per parameter sample

$$\mu_{a,t+1}(\theta_{t+1}^{(m)}) = \mu_a(x_{t+1}, \theta_{t+1}^{(m)}) \tag{19}$$

Compute sufficient statistics

$$\begin{aligned} \hat{p}_{a,t+1} &= \frac{1}{M} \sum_{m=1}^{M} I_a\left(\theta_{t+1}^{(m)}\right) \\ \hat{\sigma}_{a,t+1}^2 &= \frac{1}{M} \sum_{m=1}^{M} \left(I_a\left(\theta_{t+1}^{(m)}\right) - \hat{p}_{a,t+1}\right)^2 \end{aligned} \tag{20}$$

end for

Compute estimates

$$\begin{cases} \hat{a}_{t+1}^* = \text{argmax}_a\, \hat{p}_{a,t+1}\ , \\ \hat{p}_{a,t+1}^* = \max_a \hat{p}_{a,t+1}\ . \end{cases} \tag{21}$$

for $a = 1, \cdots, A$ **do**

Compute probability of arm not being optimal

$$\hat{p}_{FA}^{(a)} = Pr\left(\hat{p}_{a,t+1} > \hat{p}_{a,t+1}^*\right) \tag{22}$$

end for

Compute the number of candidate arm samples

$$N_{t+1} \propto \log\left(\frac{1}{\hat{p}_{FA}}\right), \quad \hat{p}_{FA} = \frac{1}{A-1} \sum_{a \neq \hat{a}_{t+1}^*} \hat{p}_{FA}^{(a)} \tag{23}$$

Draw N_{t+1} candidate arm samples

$$\hat{a}_{t+1}^{(n)} \sim \text{Cat}\left(\hat{p}_{a,t+1}\right), \quad n = 1, \cdots, N_{t+1} \tag{24}$$

Play arm

$$a_{t+1} = \text{Mode}\left(\hat{a}_{t+1}^{(n)}\right), \quad n = 1, \cdots, N_{t+1} \tag{25}$$

Observe reward y_{t+1}

Update $D = D \cup \{x_{t+1}, a_{t+1}, y_{t+1}\}$

end for

The proposed sampling policy reduces to a probabilistic matching regime when uncertain about the arms, (i.e., $N_t \approx 1$), but favors exploitation ($N_t \gg 1$) when the probability of picking a suboptimal arm is low. In other words, double sampling exploits only when confident about the learned probabilities ($\hat{\sigma}_{a,t+1} \rightarrow 0, N_t \gg 1$), and picks the arm with the highest probability $\hat{p}_{a,t+1}$. However, for $N_t \approx 1$, a randomized probability matching is in play. Note that Thompson sampling is a special case of double sampling, when $N_t = 1 \ , \forall t$.

To conclude, note that the sampling-based policy decides on the action to take next, by drawing from an approximation to the posterior density $p_{a,t+1}$. Precisely, by probability matching the expected return of each arm, which is estimated via Monte Carlo as in Eqn. (11). For the derivation of performance bounds in multi-armed bandit problems and, in particular, regret bounds for posterior sampling techniques, one studies the expected returns of the arms. Due to Monte Carlo guarantees on the convergence of the computed estimates ($\lim_{M \rightarrow \infty} \hat{p}_{a,t+1} = p_{a,t+1}$), and the random probability matching nature of double sampling, the regret bounds for our proposed technique are of the same order as those of any posterior sampling technique (3, 20). We argue that the discrepancies are on the multiplicative constants, which we evaluate in the following section.

4 Evaluation

We now empirically evaluate the performance of double sampling in both discrete and continuous contextual multi-armed bandit settings. We compare the performance of our proposed algorithm, to that of Thompson sampling (9) and Bayes-UCB (11).

On the one hand, (9) show empirically the significant advantages Thompson sampling offers for the Bernoulli and other cases, while theoretical guarantees are provided in (2, 3, 12, 19, 20). On the other, (11) prove the asymptotic optimality of Bayes-UCB's finite-time regret bound for the Bernoulli case, and argue that it provides an unifying framework for several variants of the UCB algorithm for different bandit problems: parametric multi-armed bandits and linear Gaussian bandits.

We compare double sampling as in Algorithm 1 to these two state-of-the-art algorithms, in order to provide empirical evidence of the reduced cumulative regret of our proposed approach. We define cumulative regret as

$$R_t = \sum_{\tau=0}^{t} \mathbb{E}\left\{(y_\tau^* - y_\tau)\right\} = \sum_{\tau=0}^{t} \mu_\tau^* - \bar{y}_\tau \ , \tag{26}$$

where for each time instant t, μ_t^* denotes the expected reward of the optimal arm and $\bar{y}_t$ the empirical mean of the observed rewards under the executed policy. Note that even if the bandits considered are stationary (i.e., parameters are not dynamic), the expected rewards are indexed with time to accommodate their dependency with potentially time-dependent contexts x_t.

4.1 Bernoulli bandits

Bernoulli bandits are well suited for applications with binary rewards (i.e., success or failure of an action). The rewards of each arm are modeled as independent draws from a Bernoulli distribution with success probabilities θ_a, i.e.,

$$f_a(y|\theta) = \theta_a^y (1 - \theta_a)^{(1-y)} \ . \tag{27}$$

For a Bernoulli reward distribution, the posterior parameter update can be computed using the conjugate prior distribution $f(\theta_a|\alpha_{a,0}, \beta_{a,0}) = \text{Beta}\,(\theta_a|\alpha_{a,0}, \beta_{a,0})$. After observing actions $a_{1:t}$ and rewards $y_{1:t}$, the posterior parameter distribution follows an updated Beta distribution

$$f(\theta_a|a_{1:t}, y_{1:t}, \alpha_{a,0}, \beta_{a,0}) = f(\theta_a|\alpha_{a,t}, \beta_{a,t}) = \text{Beta}\,(\theta_a|\alpha_{a,t}, \beta_{a,t}) \ , \tag{28}$$

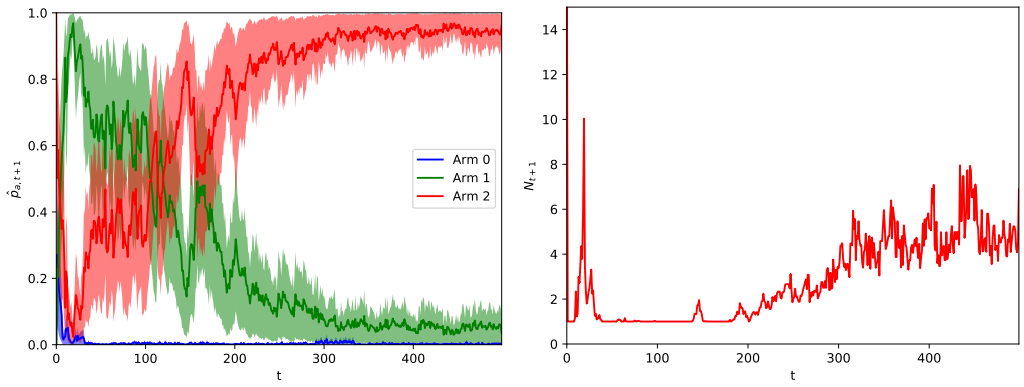
with sequential updates

$$\begin{cases} \alpha_{a,t} = \alpha_{a,t-1} + y_t \cdot \mathbb{1}[a_t = a] \ , \\ \beta_{a,t} = \beta_{a,t-1} + (1 - y_t) \cdot \mathbb{1}[a_t = a] \ , \end{cases} \tag{29}$$

or, alternatively, batch updates of the following form

$$\begin{cases} \alpha_{a,t} = \alpha_{a,0} + \sum_{t|a_t=a} y_t \ , \\ \beta_{a,t} = \beta_{a,0} + \sum_{t|a_t=a} (1 - y_t) \ . \end{cases} \tag{30}$$

The sequential Bayesian learning process for a three-armed Bernoulli bandit with parameters $\theta = (0.4\ 0.7\ 0.8)$ is illustrated in Fig. 1a. We show the evolution of the probability of each arm being optimal as computed by our proposed algorithm: i.e., the Monte Carlo approximation to $p_{a,t+1}$. For all results to follow, we use $M = 1000$ Monte Carlo samples, as larger Ms do not significantly improve regret performance. In Fig. 1b, we illustrate how double sampling is *automatically* adjusted according to the uncertainty of the learning process, via the number of arm samples to draw (i.e., N_{t+1} in Eqn. (15)).



(a) $\hat{p}_{a,t+1}$ over time. (b) N_{t+1} over time.

Figure 1: Illustrative execution of double sampling.

Let us elaborate on the exploration-exploitation tradeoff by following the double sampling bandit execution shown in Fig. 1. Observe how arm 0 is quickly discarded as a good candidate, while the decision over which of the other two arms is optimal requires further learning. For some time ($t < 200$), there is high uncertainty about the properties of these two arms (high variance in Fig. 1a). Thus, double sampling favors exploration ($N_{t+1} \approx 1$ in Fig. 1b), until the uncertainty about which arm is best is reduced. Once the algorithm becomes more certain about the better reward properties of arm 2 ($t > 200$), double sampling gradually favors a greedier policy ($N_{t+1} > 1$).

All in all, within periods of high uncertainty, the number of samples N_{t+1} is kept low (i.e., exploration); on the contrary, when the learning is more accurate, it increases (i.e., exploitation). By means of the double sampling technique, we account for the uncertainty in the learning process and thus, the proposed algorithm can reduce the variance over the actions taken.

We plot in Fig. 2 the empirical probabilities of each algorithm playing the optimal arm over 5000 realizations[1] of a Bernoulli bandit with $A = 2, \theta = (0.4\ 0.8)$. Observe how, even if in expectation all algorithms take the same actions, the action variability of double sampling is considerably smaller.

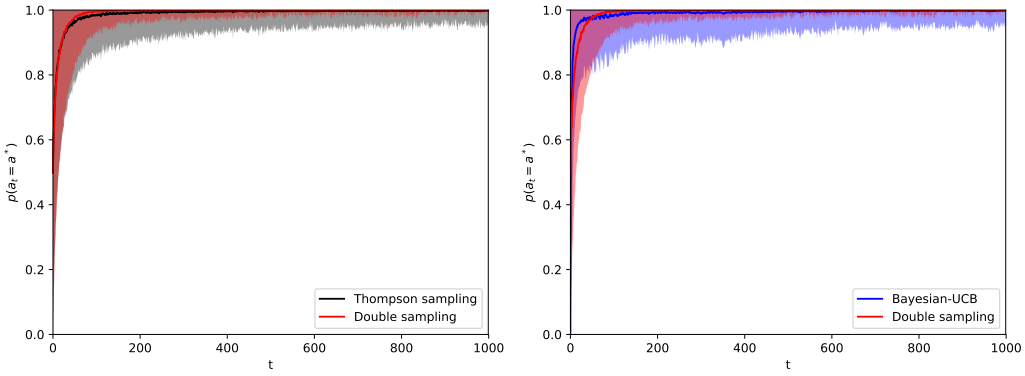


Figure 2: Averaged correct action probability (standard deviation as shaded region) with $A = 2, \theta = (0.4\ 0.8)$.

As a result, the cumulative regret of our proposed technique is lower than those of the compared alternatives, i.e., Thompson sampling and Bayes-UCB, (see averaged cumulative regrets in Fig. 3).

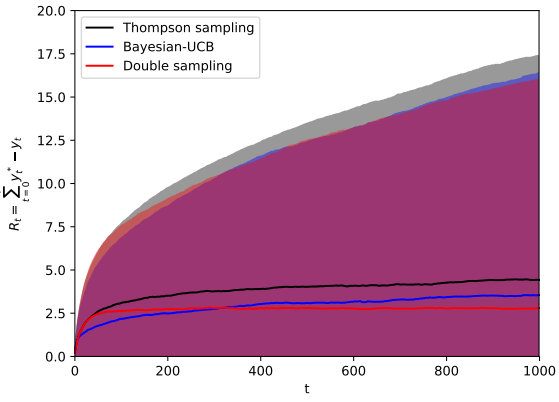


Figure 3: Averaged cumulative regret (standard deviation shown as shaded region) with $A = 2, \theta = (0.4\ 0.8)$.

[1] All averaged results in this work are computed over 5000 realizations of the same set of parameters.

For any bandit problem, the difficulty of learning the properties of each arm is a key factor on its regret performance. Intuitively, this difficulty relates to how close the expected rewards of each arm are. Mathematically, it can be captured by the divergence between arm reward distributions. By computing the minimum Kullback-Leibler (KL) divergence between arms, one quantifies how "difficult" a multi-armed bandit problem is, as established by the regret lower-bound in (14).

We evaluate the relative difference between the averaged cumulative regret of our proposed double sampling technique and the alternatives, i.e.,

$$\Delta_t^{(TS)} = \frac{R_t^{(DS)}}{R_t^{(TS)}} - 1 \quad \text{and} \quad \Delta_t^{(B-UCB)} = \frac{R_t^{(DS)}}{R_t^{(B-UCB)}} - 1 \,, \tag{31}$$

where $R_t^{(DS)}$ denotes the regret of the proposed double sampling approach at time t, $R_t^{(TS)}$, that of Thompson sampling, and $R_t^{(B-UCB)}$, that of Bayes-UCB.

We show in Fig. 4 results for the above metric indexed by the KL divergence of a wide range of Bernoulli bandit parameterizations[2]. Note that the KL metric may map many parameter combinations to the same point in Fig. 4.

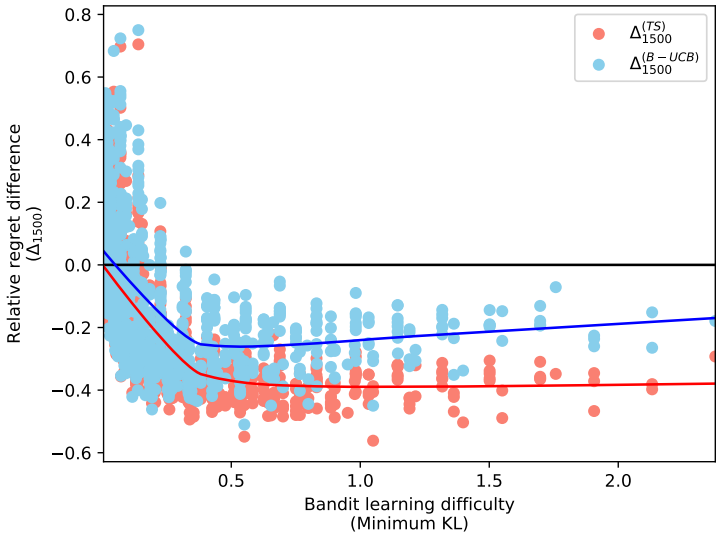


Figure 4: Relative average cumulative regret differences at $t = 1500$.

Double sampling performs significantly better than the alternatives when it is certain about the learned arm parameters. We obtain cumulative regret reductions around 25% and 40% when compared to B-UCB (with optimal quantile parameter $\alpha_t = 1/t$ as in (11)) and Thompson sampling, respectively. However, when the best arms are very similar ($KL < 0.25$), performance worsens. First, recall that the regret lower-bound increases for all bandits with small KL values (14). Second, note that when the properties of the arms are very similar to each other, our algorithm resorts to a Thompson sampling-like policy ($N_{t+1} \approx 1$), yielding near-equivalent performance.

[2]Bernoulli bandits with $A = 2$ and $A = 3$ arms, for all per-arm parameter permutations in the range $\theta_a \in [0, 1]$ with grid size 0.05.

Finally, we observe that for the challenging cases (i.e., $KL < 0.25$) cumulative regret shows high variance for the three alternatives (points are scattered in Fig. 4). On the contrary, when the difference between arm properties is distinguishable ($KL > 0.25$), the proposed double sampling technique considerably reduces cumulative regret for Bernoulli multi-armed bandits.

4.2 Contextual linear Gaussian bandits

Another set of well studied bandits are those with continuous reward functions and, in particular, those with contextual dependencies. That is, the reward distribution of each arm is dependent on a time-varying d-dimensional context vector $x_t \in \mathbb{R}^d$.

The contextual linear Gaussian bandit model is suited for these scenarios, where the expected reward of each arm is linearly dependent on the context $x \in \mathbb{R}^d$, and the idiosyncratic parameters of the bandit $\theta \equiv \{w, \sigma\}$. That is, the per-arm reward distribution follows

$$f_a(y|x,\theta) = \mathcal{N}\left(y|x^\top w_a, \sigma_a^2\right) \ . \tag{32}$$

For such reward distribution, the posterior can be computed with the Normal Inverse Gamma conjugate prior distribution

$$\begin{aligned} f(w_a, \sigma_a^2|u_{a,0}, V_{a,0}, \alpha_{a,0}, \beta_{a,0}) &= \text{NIG}\left(w_a, \sigma_a^2|u_{a,0}, V_{a,0}, \alpha_{a,0}, \beta_{a,0}\right) \\ &= \mathcal{N}\left(w_a|u_{a,0}, \sigma_a^2 V_{a,0}\right) \cdot \Gamma^{-1}\left(\sigma_a^2|\alpha_{a,0}, \beta_{a,0}\right) \ . \end{aligned} \tag{33}$$

Given previous actions $a_{1:t}$, contexts $x_{1:t}$ and rewards $y_{1:t}$, one obtains the following posterior

$$f(w_a, \sigma_a^2|a_{1:t}, y_{1:t}, u_{a,0}, V_{a,0}, \alpha_{a,0}, \beta_{a,0}) = \text{NIG}\left(w_a, \sigma_a^2|u_{a,t}, V_{a,t}, \alpha_{a,t}, \beta_{a,t}\right) \ , \tag{34}$$

where the parameters of the posterior are sequentially updated as

$$\begin{cases} V_{a,t}^{-1} = V_{a,t-1}^{-1} + x_t x_t^\top \cdot \mathbb{1}[a_t = a] \ , \\ u_{a,t} = V_{a,t}\left(V_{a,t-1}^{-1} u_{a,t-1} + x_t y_t \cdot \mathbb{1}[a_t = a]\right) \ , \\ \alpha_{a,t} = \alpha_{a,t-1} + \frac{\mathbb{1}[a_t = a]}{2} \ , \\ \beta_{a,t} = \beta_{a,t-1} + \frac{\mathbb{1}[a_t = a](y_{t_a} - x_t^\top \theta_{a,t-1})^2}{2\left(1 + x_t^\top \Sigma_{a,t-1} x_t\right)} \ . \end{cases} \tag{35}$$

Alternatively, if data is collected in batches, one updates the posterior with

$$\begin{cases} V_{a,t}^{-1} = V_{a,0}^{-1} + x_{1:t|t_a} x_{1:t|t_a}^\top \ , \\ u_{a,t} = V_{a,t}\left(V_{a,0}^{-1} u_{a,0} + x_{1:t|t_a} y_{1:t|t_a}\right) \ , \\ \alpha_{a,t} = \alpha_{a,0} + \frac{|t_a|}{2} \ , \\ \beta_{a,t} = \beta_{a,0} + \frac{\left(y_{1:t|t_a}^\top y_{1:t|t_a} + u_{a,0}^\top V_{a,0}^{-1} u_{a,0} - u_{a,t}^\top V_{a,t}^{-1} u_{a,t}\right)}{2} \ , \end{cases} \tag{36}$$

where $t_a = \{t|a_t = a\}$ indicates the set of time instances when arm a is played.

We evaluate double sampling for the multi-armed contextual Gaussian bandit with uniform and uncorrelated context, i.e., $x_{i,t} \sim \mathcal{U}(0,1), i \in \{1, \cdots, d\}, t \in \mathbb{N}$.

We again use the minimum KL divergence as a proxy for bandit complexity. The divergence is model agnostic, as many parameter combinations for any model may map to the same KL divergence value.

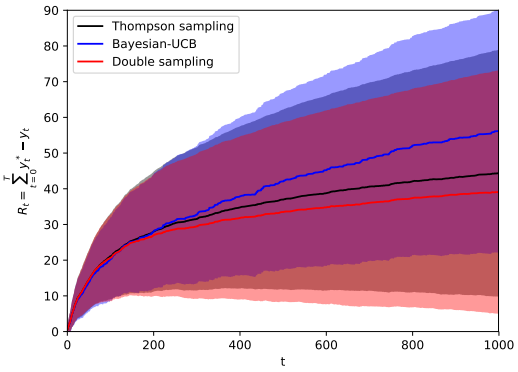


Figure 5: Averaged cumulative regret comparison (standard deviation shown as shaded region) with $A = 2$, $w_0 = (0.4\ 0.4)^\top$, $w_1 = (0.8\ 0.8)^\top$, $\sigma_0 = \sigma_1 = 0.2$.

We provide results for a specific two-armed contextual Gaussian bandit in Fig. 5, and in Fig. 6, average cumulative regret relative differences (as in Eqn. (31)) for a wide range of parameterizations[3] of two-dimensional contextual linear Gaussian bandits with two arms.

Again, when the reward difference between arms is easy to learn (KL>0.25), double sampling attains significant cumulative regret reductions. The regret improvement is most evident for models with significant KL divergence, with cumulative regret reductions of up to 40% and 50% when compared to Thompson sampling and Bayes-UCB, respectively.

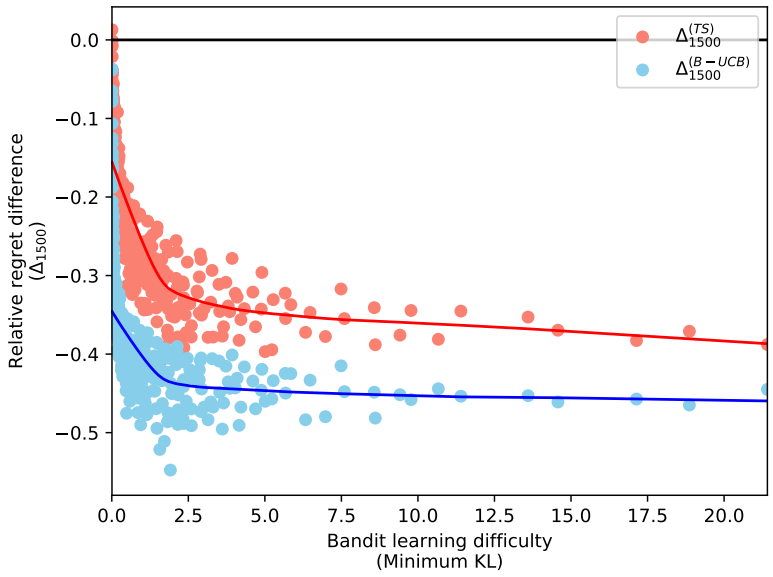


Figure 6: Relative average cumulative regret differences at $t = 1500$.

[3]Contextual linear Gaussian bandits with per-dimension parameter $w_i \in [-1, 1], i \in \{1, 2\}$ with gaps of 0.1, and $\sigma \in [0.1, 1]$ with step size of 0.1.

We argue that the comparative performance loss of Bayes-UCB in the contextual Gaussian case relates to the $\alpha_t = 1/t$ quantile value proposed by (11). Its justification comes from the bounds established for the Bernoulli bandit case, but there are no guarantees provided for other bandits. That is, the optimal quantile values for Bayes-UCB are problem dependent, and require careful analytical derivations. On the contrary, our proposed double sampling algorithm does not require any manual tuning, as it autonomously balances the exploration-exploitation tradeoff by adjusting the number of candidate arm samples N_{t+1} based on the learning uncertainty.

5 Conclusion

We have presented a new sampling-based probability matching technique for the multi-armed bandit setting. We formulated the problem as a Bayesian sequential learning one, and leveraged random sampling to overcome two of its main challenges: approximating the analytically unsolvable integrals, and automatically balancing the exploration-exploitation tradeoff. We empirically show that additional sampling from the model, which is in many application domains inexpensive in comparison with interacting with the world, can provide improved statistics and, ultimately, reduced regrets. Encouraged by these findings, we aim at implementing this technique with other reward distributions and extending it to real application datasets.

5.1 Software and Data

The implementation of the proposed method is available in this public repository. It contains all the software required for replication of the findings of this study.

Acknowledgments

This research was supported in part by NSF grant SCH-1344668. We thank Hang Su and Edward Peng Yu for their feedback and comments on this work.

References

[1] S. Agrawal and N. Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. *CoRR*, abs/1111.1797, 2011.

[2] S. Agrawal and N. Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. *CoRR*, abs/1209.3352, 2012.

[3] S. Agrawal and N. Goyal. Further Optimal Regret Bounds for Thompson Sampling. *CoRR*, abs/1209.3353, 2012.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2-3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352.

[5] R. Bellman. A Problem in the Sequential Design of Experiments. *Sankhya: The Indian Journal of Statistics (1933 - 1960)*, 16(3/4):221–229, 1956.

[6] J. M. Bernardo and A. F. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, 2009. ISBN 9780470317716. doi: 10.1002/9780470316870.

[7] M. Brezzi and T. L. Lai. Incomplete Learning from Endogenous Data in Dynamic Allocation. *Econometrica*, 68(6):1511–1516, 2000. ISSN 1468 0262. doi: 10.1111/1468-0262.00170.

[8] M. Brezzi and T. L. Lai. Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics and Control*, 27(1):87 – 108, 2002. ISSN 0165-1889. doi: https://doi.org/10.1016/S0165-1889(01)00028-8.

[9] O. Chapelle and L. Li. An Empirical Evaluation of Thompson Sampling. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2249–2257. Curran Associates, Inc., 2011.

[10] J. C. Gittins. Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979. ISSN 00359246.

[11] E. Kaufmann, O. Cappe, and A. Garivier. On Bayesian Upper Confidence Bounds for Bandit Problems. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 592–600, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[12] N. Korda, E. Kaufmann, and R. Munos. Thompson Sampling for 1-Dimensional Exponential Family Bandits. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1448–1456. Curran Associates, Inc., 2013.

[13] T. L. Lai. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. *The Annals of Statistics*, 15(3):1091–1114, 1987. ISSN 00905364.

[14] T. L. Lai and H. Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6(1):4–22, mar 1985. ISSN 0196-8858. doi: 10.1016/0196-8858(85)90002-8.

[15] L. Li, W. Chu, J. Langford, and R. E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. *CoRR*, abs/1003.0146, 2010.

[16] O.-A. Maillard, R. Munos, and G. Stoltz. Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences. In *Conference On Learning Theory*, 2011.

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

[18] H. Robbins. A sequential decision procedure with a finite memory. *Proceedings of the National Academy of Science*, (42):920 – 923, 1956.

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

[20] D. Russo and B. V. Roy. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

[21] S. L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010. ISSN 1526-4025. doi: 10.1002/asmb.874.

[22] S. L. Scott. Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, 31:37–49, 2015. Special issue on actual impact and future perspectives on stochastic modelling in business and industry.

[23] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press: Cambridge, MA, 1998.

[24] W. 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. ISSN 00063444.

[25] W. R. Thompson. On the Theory of Apportionment. *American Journal of Mathematics*, 57(2):450–456, 1935. ISSN 00029327, 10806377.