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

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

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

August 22, 2017

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 works. One general class of algorithms for such learning is the multi-armed bandit setting (in which sequential interactions are independent and identically distributed) and the related contextual bandit case, in which the distribution depends on different information or 'context' presented with each interaction. Thompson sampling, though introduced in the 1930s, has recently been shown to perform well and to enjoy provable optimality properties, while at the same time permitting generative, interpretable modeling. In a Bayesian setting, prior knowledge is incorporated and the computed posteriors naturally capture the full state of knowledge. 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. 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. We empirically show that it out-performs (in the sense of regret) Thompson sampling in two classical illustrative cases, i.e., the multi-armed bandit problem with and without context.

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(1952), Robbins(1956)]. Since then, 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 ([Auer et al.(2002)]). In the latter case, [Gittins(1979)] formulated a method based on computing the optimal strategy for some 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 ([Brezzi and Lai(2002)]). These and other intrinsic challenges of the method have limited its applicability ([Sutton and Barto(1998)]).

[Lai and Robbins(1985)] and [Lai(1987)] introduced another class of algorithms, based on upper confidence intervals of the expected reward of each arm, for which strong theoretical guarantees were proved. Nevertheless, these algorithms might be far from optimal in the presence of dependent and more general reward distributions ([Scott(2010)]).

More recently, the problem has re-emerged both from a practical (importance in e-commerce and web applications, e.g. [Li et al.(2010)]) and a theoretical (research on probability matching algorithms and their regret bounds, e.g. [Agrawal and Goyal(2011)] and [Maillard et al.(2011)]) point of view. Contributing to this revival was the observation that one of the oldest heuristics to address the exploration-exploitation trade-off, i.e., [Thompson(1935), Thompson(1933)] sampling, has been empirically proven to perform satisfactorily (see [Chapelle and Li(2011)] and [Scott(2015)] for details). Contemporaneously, theoretical study established several performance bounds, both for problems with and without context ([Agrawal and Goyal(2012a), Agrawal and Goyal(2012b), Korda et al.(2013), Russo and Roy(2014), Russo and Roy(2016)]).

In this work, we are interested in the randomized probability matching approach, as it connects to the Bayesian learning paradigm which readily facilitates modeling and algorithm development in both sequential and batch processing scenarios. Specifically, we establish how one can extract more information about the environment by casting the problem as a Bayesian sequential learning process, so that better informed decisions can be made, leading to a lower regret than in the Thompson sampling approach. We do not make any distributional assumption, as long as Bayesian posterior updates are computable, and consider complicated relationships among action rewards. 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.

We propose a technique for the multi-armed bandit problem that is based on Monte Carlo sampling (to approximate otherwise unsolvable integrals) and 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 each arm, exploring otherwise. We empirically show that the proposed algorithm provides improved average performance, with significant 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 approach 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, and similarly $y_{1:t} \equiv (y_1, \cdots, y_t)$ refers to the sequence of observed rewards up to time t.

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 most common metric used. 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 ([Bellman(1956), Gittins(1979)]), and it fails to generalize to more realistic reward distributions and scenarios ([Scott(2010)]).

The biggest challenge occurs when the parameters are unknown, as one might end up playing the wrong arm forever if incomplete learning occurs ([Brezzi and Lai(2000)]). Practitioners have often turned to heuristics to overcome these issues, and amongst them, the randomized probability matching, i.e., playing each arm in proportion to its probability of being optimal, is a particularly appealing one.

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,

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

Note that there is an inherent uncertainty about the unknown properties of the arms, as Eqn. 4 is parameterized by θ, while Eqn. 3 is not. In order to compute a solution to this problem, recasting it as a Bayesian learning problem is of great help, as it allows for direct connection with the randomized probability matching technique ([Scott(2010)]). The randomized probability matching approach has shown to be easy to implement, efficient and broadly applicable.

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 parameters 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 ([Bernardo and Smith(2009)]).

However, the solution to the probability matching equation 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 the algorithm in 1.

3.1 The Bayesian multi-armed bandit

We are interested in computing, after playing arms $a_{1:t}$ and observing rewards $y_{1:t}$, the probability of each arm a being optimal for the next time instant, i.e.,

$$p_{a,t+1} \equiv \Pr\left[a = a^*_{t+1} \middle| a_{1:t}, y_{1:t}\right] = \Pr\left[\mu_a(\theta) = \max\{\mu_1(\theta), \cdots, \mu_A(\theta)\} \middle| a_{1:t}, y_{1:t}\right]. \tag{5}$$

In practice, one needs to account for the lack of knowledge of each arm's properties, i.e., the unknown parameters θ. 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 t,

$$\begin{aligned} p_{a,t+1} \equiv \Pr\left[a = a^*_{t+1} \middle| a_{1:t}, y_{1:t}\right] &= f(a = a^*_{t+1} | a_{1:t}, y_{1:t}) \\ &= \int f(a | a_{1:t}, y_{1:t}, \theta) f(\theta | a_{1:t}, y_{1:t}) \mathrm{d}\theta \,. \end{aligned} \tag{6}$$

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

$$f(\theta | a_{1:t}, y_{1:t}) \propto f_{a_t}(y_t|\theta) f(\theta | a_{1:t-1}, y_{1:t-1}) = \left[\prod_{\tau=1}^{t} f_{a_\tau}(y_\tau|\theta)\right] f(\theta) \,. \tag{7}$$

This posterior provides information (with uncertainty) about the state 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

$$\begin{aligned} p_{a,t+1} \equiv f(a = a^*_{t+1}|a_{1:t}, y_{1:t}) &= \int f(a|a_{1:t}, y_{1:t}, \theta) f(\theta|a_{1:t}, y_{1:t}) \mathrm{d}\theta \\ &= \int I_a(\theta) f(\theta|a_{1:t}, y_{1:t}) \mathrm{d}\theta \ . \end{aligned} \tag{8}$$

3.2 Monte-Carlo integration

We harness the power of Monte Carlo sampling to compute the otherwise analytically intractable integral in Eqn. 8.

We obtain a Monte Carlo approximation to $p_{a,t+1} \in [0, 1]$ as follows:

1. Draw M parameter samples from the updated posterior distribution

$$\theta^{(m)} \sim f(\theta|a_{1:t}, y_{1:t}), \ m = \{1, \cdots, M\} \ . \tag{9}$$

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

$$a^*_{t+1}(\theta^{(m)}) = \underset{a}{\operatorname{argmax}} \, \mu_a(\theta^{(m)}) \ . \tag{10}$$

3. Define the random measure approximation to the probability distribution in Eqn. 8 as

$$f_M(a = a^*_{t+1}|a_{1:t}, y_{1:t}) = \frac{1}{M} \sum_{m=1}^{M} \delta \left(a - a^*_{t+1}(\theta^{(m)}) \right) \ , \tag{11}$$

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

4. Estimate the first- and second-order sufficient statistics of the distribution

$$\begin{cases} \hat{p}_{a,t+1} = \frac{1}{M} \sum_{m=1}^{M} I_a \left(\theta^{(m)} \right) \ , \\ \hat{\sigma}^2_{a,t+1} = \frac{1}{M} \sum_{m=1}^{M} \left(I_a \left(\theta^{(m)} \right) - \hat{p}_{a,t+1} \right)^2 \ . \end{cases} \tag{12}$$

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

$$\begin{cases} \hat{a}^*_{t+1} = \operatorname{argmax}_a \hat{p}_{a,t+1} \ , \\ \hat{p}^*_{a,t+1} = \max_a \hat{p}_{a,t+1} \ . \end{cases} \tag{13}$$

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. 8, and leverage the estimated sufficient statistics in Eqn. 12 to balance the exploration-exploitation tradeoff.

We draw candidate arm samples from the random measure in Eqn. 11, 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., $\hat{\sigma}^2_{a,t+1}$). 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

$$p^{(a)}_{FA} = Pr\left(p_{a,t+1} > p^*_{a,t+1}\right) = 1 - F_{p_{a,t+1}}(p^*_{a,t+1}) \;, \tag{14}$$

where $p^*_{a,t+1} = \max_a p_{a,t+1}$. Since we can not exactly evaluate $p_{a,t+1}$, we resort to our Monte Carlo estimates in Eqn. 12. The true cumulative density function $F_{p_{a,t+1}}(\cdot)$ is analytically intractable as well, but we approximate it with a truncated Gaussian

$$F_{p_{a,t+1}}(p^*_{a,t+1}) \approx \Phi_{[0,1]}\left(\frac{\hat{p}_{a,t+1} - \hat{p}^*_{a,t+1}}{\hat{\sigma}_{a,t+1}}\right) \;. \tag{15}$$

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

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

$$\begin{aligned} N_{t+1} &\propto \log\left(\frac{1}{p_{FA}}\right), \\ \text{with } \; p_{FA} &= \frac{1}{K-1}\sum_{a \neq \hat{a}^*_{t+1}} p^{(a)}_{FA} \;, \qquad p^{(a)}_{FA} \approx 1 - \Phi_{[0,1]}\left(\frac{\hat{p}_{a,t+1} - \hat{p}^*_{a,t+1}}{\hat{\sigma}_{a,t+1}}\right) \;. \end{aligned} \tag{16}$$

2. Draw N_{t+1} candidate arm samples from the random measure in Eqn. 11

$$\hat{a}^{(n)}_{t+1} \sim \text{Cat}\left(\hat{p}_{a,t+1}\right) \;, \qquad n = 1, \cdots, N_{t+1} \;. \tag{17}$$

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

$$a_{t+1} = \text{Mode}\left(\hat{a}^{(n)}_{t+1}\right) \;, \qquad n = 1, \cdots, N_{t+1} \;. \tag{18}$$

By allowing for N_{t+1} to be adjusted based upon the uncertainty of the learning process, we balance the exploration-exploitation tradeoff. 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 ($\hat{\sigma}_{a,t+1} \to 0, N_t \gg 1$), 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 $N_t = 1 \;, \forall t$. We present full details of the proposed double sampling algorithm in 1.

We conclude by noting that the double sampling policy decides on the action to take at every time instant 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 Equation 12. The expected returns are key components for the derivation

of performance bounds in multi-armed bandit problems and, in particular, the regret bounds for Thompson sampling. Due to the random probability matching nature of double sampling, the regret bounds for our proposed technique are of the same order as those of Thompson sampling, with differences on the multiplicative constants. We empirically evaluate this gap in the following section.

4 Evaluation: Thompson sampling vs. double sampling

We now provide empirical evidence of our double sampling's lower regret relative to Thompson sampling. We consider both discrete and continuous multi-armed bandits, and the contextual setting as well. We compare the performance of our algorithm, as presented in 1, to that of Thompson sampling, for both the Bernoulli and the contextual linear Gaussian bandits. These are well-studied multi-armed bandits, with theoretical and experimental results in the literature, where it has been extensively proven that Thompson sampling offers significant advantages over other popular approaches.

As is standard in the literature, we measure performance in the cumulative regret sense, i.e.,

$$R_t = \sum_{\tau=0}^{t} \mathbb{E}\left\{(y_\tau^* - y_\tau)\right\} = \sum_{\tau=0}^{t} \mu_\tau^* - \bar{y}_\tau \ , \tag{27}$$

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.

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.,

$$f_a(y|\theta) = \theta_a^y (1-\theta_a)^{(1-y)} \ . \tag{28}$$

For this 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

$$f(\theta_a|a_{1:t}, y_{1:t}, \alpha_{a,0}, \beta_{a,0}) = f(\theta_a|\alpha_{a,t}, \beta_{a,t}) = \text{Beta}(\theta_a|\alpha_{a,t}, \beta_{a,t}) \ , \tag{29}$$

with sequential updates

$$\begin{cases} \alpha_{a,t} = \alpha_{a,t-1} + y_t \cdot \mathbb{1}[a_t = a] \ , \\ \beta_{a,t} = \beta_{a,t-1} + (1-y_t) \cdot \mathbb{1}[a_t = a] \ , \end{cases} \tag{30}$$

or, alternatively, batch updates of the following form

$$\begin{cases} \alpha_{a,t} = \alpha_{a,0} + \sum_{t|a_t=a} y_t \ , \\ \beta_{a,t} = \beta_{a,0} + \sum_{t|a_t=a} (1-y_t) \ . \end{cases} \tag{31}$$

We first introduce and dissect the key aspects of double sampling in a particular realization of a Bernoulli bandit in Fig. 1, before providing extensive evaluation results in Figs. 2 and 3.

Algorithm 1 Bayesian Double Sampling algorithm

Require: Number of arms A
Require: Prior over model parameters $f(\theta)$
Require: Per-arm reward distributions $f_a(y|\theta)$
Require: Horizon T
$D = \emptyset$
for $t = 1, \cdots, T$ **do**
Draw M parameter samples from the updated posterior distribution

$$\theta_{t+1}^{(m)} \sim f(\theta|a_{1:t}, y_{1:t}),\ m = \{1, \cdots, M\}\ . \tag{19}$$

If applicable, receive context x_{t+1}
for $a = 1, \cdots, A$ **do**
Compute the expected reward for each parameter sample $\theta_{t+1}^{(m)}$

$$\mu_{a,t+1}(\theta_{t+1}^{(m)}) = \mu_a(x_{t+1}, \theta_{t+1}^{(m)}) = \mathbb{E}_a\{y|x_{t+1}, \theta_{t+1}^{(m)}\} \tag{20}$$

Compute the first- and second-order sufficient statistics of each arm being optimal

$$\begin{cases} \hat{p}_{a,t+1} = \frac{1}{M}\sum_{m=1}^{M} I_a\left(\theta_{t+1}^{(m)}\right)\ , \\ \hat{\sigma}_{a,t+1}^2 = \frac{1}{M}\sum_{m=1}^{M}\left(I_a\left(\theta_{t+1}^{(m)}\right) - \hat{p}_{a,t+1}\right)^2\ , \\ \text{where } I_a(\theta_{t+1}^{(m)}) = \begin{cases} 1,\ \mu_a(\theta_{t+1}^{(m)}) = \max\{\mu_1(\theta_{t+1}^{(m)}), \cdots, \mu_A(\theta_{t+1}^{(m)})\}\ , \\ 0,\ \text{otherwise}\ . \end{cases} \end{cases} \tag{21}$$

end for
Estimate the optimal arm and its probability of being optimal

$$\begin{cases} \hat{a}_{t+1}^* = \text{argmax}_a\, \hat{p}_{a,t+1}\ , \\ \hat{p}_{a,t+1}^* = \max_a \hat{p}_{a,t+1}\ . \end{cases} \tag{22}$$

for $a = 1, \cdots, A$ **do**
Compute the estimated probability of each arm not being the optimal arm

$$\begin{aligned} \hat{p}_{FA}^{(a)} = Pr\left(\hat{p}_{a,t+1} > \hat{p}_{a,t+1}^*\right) = 1 - F_{\hat{p}_{a,t+1}}(\hat{p}_{a,t+1}^*)\ , \\ F_{\hat{p}_{a,t+1}}(p_{a,t+1}^*) \approx \Phi_{[0,1]}\left(\frac{\hat{p}_{a,t+1} - \hat{p}_{a,t+1}^*}{\hat{\sigma}_{a,t+1}}\right)\ . \end{aligned} \tag{23}$$

end for
Compute N_{t+1}, the number of candidate arm samples to draw

$$N_{t+1} \propto \log\left(\frac{1}{\hat{p}_{FA}}\right),\ \hat{p}_{FA} = \frac{1}{A-1}\sum_{a \neq \hat{a}_{t+1}^*} \hat{p}_{FA}^{(a)}\ . \tag{24}$$

Draw N_{t+1} candidate arm samples

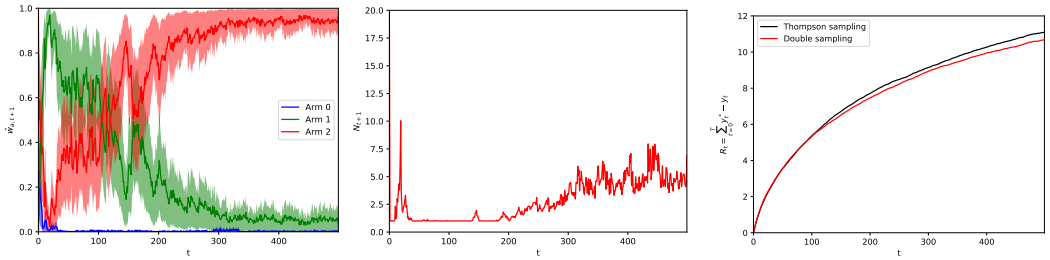
$$\hat{a}_{t+1}^{(n)} \sim \text{Cat}\left(\hat{p}_{a,t+1}\right)\ ,\qquad n = 1, \cdots, N_{t+1}\ . \tag{25}$$

Play arm

$$a_{t+1} = \text{Mode}\left(\hat{a}_{t+1}^{(n)}\right)8,\qquad n = 1, \cdots, N_{t+1}\ . \tag{26}$$

Observe reward y_{t+1}
$D = D \cup \{x_{t+1}, a_{t+1}, y_{t+1}\}$
end for

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 in Eqn. 11. For all results to follow, the Monte Carlo integration is attained with $M = 1000$ samples, as larger Ms do not significantly improve regret performance.



(a) Approximation to $w_{a,t+1}$. (b) Number of arm samples N_{t+1}. (c) Averaged cumulative regret.

Figure 1: Empirical results with double sampling.

In Fig. 1b, we illustrate how double sampling is *automatically* adjusted according to the learning process uncertainty, via the number of arm samples to draw (i.e., N_{t+1} in Eqn. 16). For periods with high uncertainty, the number of samples N_{t+1} is kept low; when the learning is more accurate, it increases.

For the case shown in Fig. 1, arm 0 (with $\theta_0 = 0.4$) is quickly discarded as being optimal, while the decision over which of the other two arms is better requires further learning. For some time ($t < 200$), there is high uncertainty about the properties of both 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.

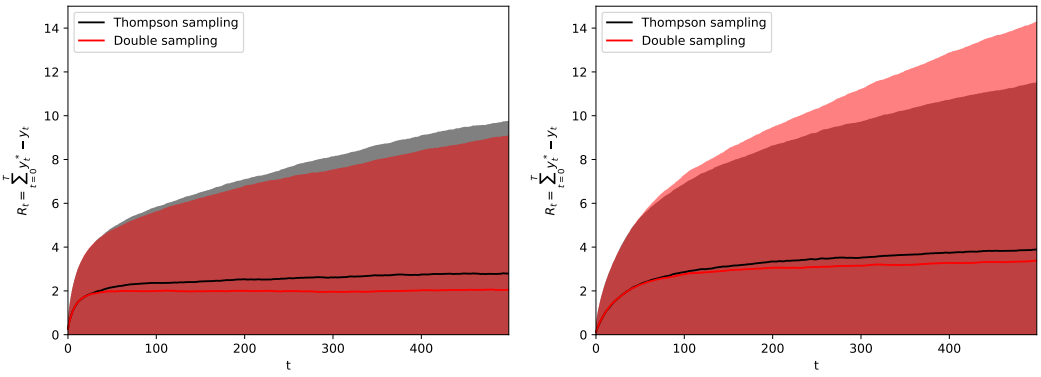
Once the algorithm becomes more certain about the better reward properties of arm 2 ($t > 200$), double sampling gradually favors a greedier policy ($N_{t+1} > 1$). As a result, the cumulative regret of our proposed technique is reduced when compared to the Thompson sampling approach (see averaged cumulative regrets in Fig. 1c). All the averaged results presented in this paper are computed over 5000 realizations of the same set of parameters.

Let us know focus on the cumulative regret of the proposed double sampling technique. In Fig. 2, we show two antagonistic examples of the performance of double sampling. Our algorithm provides a clear benefit over Thompson sampling in Fig. 2a, though it performs only slightly better than Thompson sampling in Fig. 2b. Such a cumulative regret difference is explained by the unknown nature of the bandit's arms.

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 regret. However, when the learning process is certain about the properties of the arms, double sampling can considerably reduce the method's cumulative regret. Thus, the difficulty of learning the bandit properties becomes a key evaluation criterion.

Mathematically, the bandit problem difficulty (i.e., the difference in arm properties) can be captured by the divergence between arm reward distributions. By computing the minimum Kullback-Leibler (KL) divergence between arms, one gets an insight onto how "difficult" a multi-armed bandit problem is.

We show in Fig. 3 average cumulative regret results over 5000 realizations of Bernoulli bandits with K=2 and K=3 arms, for all per-arm parameter θ_a combinations in the range



(a) Cumulative regret with $A = 2, \theta = (0.4\ 0.9)$. (b) Cumulative regret with $A = 2, \theta = (0.65\ 0.9)$.

Figure 2: Averaged cumulative regret comparison (standard deviation shown as shaded region).

$[0, 1]$ with grid size 0.05. Note that the KL metric may map many parameter combinations to the same point in Fig. 3. We evaluate the relative difference between the averaged cumulative regret of our algorithm and Thompson sampling, i.e.,

$$\Delta_t = \frac{R_t^{(DS)}}{R_t^{(TS)}} - 1 \ , \tag{32}$$

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

When the best arms are very similar (small KL), our performance is comparable to that of the Thompson sampling technique. On the contrary, double sampling performs significantly better when it is certain about the learned arm parameters (with regret reductions around 40% for many cases). In summary, Fig. 3 shows the benefit of the proposed double sampling algorithm in terms of reduced cumulative regret.

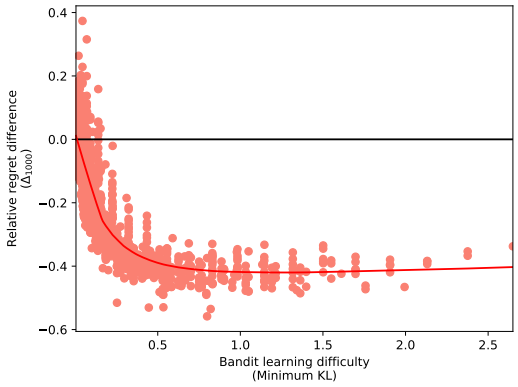


Figure 3: Relative average cumulative regret differences at $t = 1000$.

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,

$$f_a(y|x,\theta) = \mathcal{N}(y|x^\top w_a, \sigma_a^2) \ . \tag{33}$$

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

$$\begin{aligned} f(w_a, \sigma_a^2|u_{a,0}, V_{a,0}, \alpha_{a,0}, \beta_{a,0}) &= \text{NIG}(w_a, \sigma_a^2|u_{a,0}, V_{a,0}, \alpha_{a,0}, \beta_{a,0}) \\ &= \mathcal{N}(w_a|u_{a,0}, \sigma_a^2 V_{a,0}) \cdot \Gamma^{-1}(\sigma_a^2|\alpha_{a,0}, \beta_{a,0}) \ . \end{aligned} \tag{34}$$

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

$$f(w_a, \sigma_a^2|a_{1:t}, y_{1:t}, u_{a,0}, V_{a,0}, \alpha_{a,0}, \beta_{a,0}) = \text{NIG}(w_a, \sigma_a^2|u_{a,t}, V_{a,t}, \alpha_{a,t}, \beta_{a,t}) \ , \tag{35}$$

where the parameters of the posterior are sequentially updated as

$$\begin{cases} V_{a,t}^{-1} = V_{a,t-1}^{-1} + x_t x_t^\top \cdot \mathbb{1}[a_t = a] \ , \\ u_{a,t} = V_{a,t}\left(V_{a,t-1}^{-1} u_{a,t-1} + x_t y_t \cdot \mathbb{1}[a_t = a]\right) \ , \\ \alpha_{a,t} = \alpha_{a,t-1} + \frac{\mathbb{1}[a_t=a]}{2} \ , \\ \beta_{a,t} = \beta_{a,t-1} + \frac{\mathbb{1}[a_t=a](y_{t_a} - x_t^\top \theta_{a,t-1})^2}{2\left(1 + x_t^\top \Sigma_{a,t-1} x_t\right)} \ . \end{cases} \tag{36}$$

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

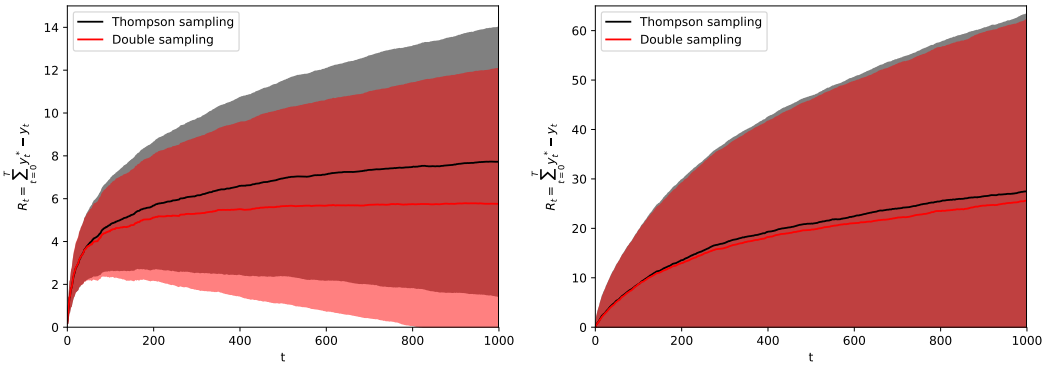
$$\begin{cases} V_{a,t}^{-1} = V_{a,0}^{-1} + x_{1:t|t_a} x_{1:t|t_a}^\top \ , \\ u_{a,t} = V_{a,t}\left(V_{a,0}^{-1} u_{a,0} + x_{1:t|t_a} y_{1:t|t_a}\right) \ , \\ \alpha_{a,t} = \alpha_{a,0} + \frac{|t_a|}{2} \ , \\ \beta_{a,t} = \beta_{a,0} + \frac{\left(y_{1:t|t_a}^\top y_{1:t|t_a} + u_{a,0}^\top V_{a,0}^{-1} u_{a,0} - u_{a,t}^\top V_{a,t}^{-1} u_{a,t}\right)}{2} \ , \end{cases} \tag{37}$$

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

We evaluate double sampling for a wide set of multi-armed contextual Gaussian bandit parameterizations, where the context is independent and identically distributed (uniform and uncorrelated, i.e., $x_{i,t} \sim \mathcal{U}(0,1), i \in \{1, \cdots, d\}, t \in \mathbb{N}$).

We provide results in Fig. 4 for two particular parameterization of the two-armed contextual Gaussian bandit. We observe that when the arms' reward distributions are comparable, our method performs similar to Thompson sampling (see Fig. 4b). However, when the difference between arms is easier to learn (as in Fig. 4a), double sampling attains reduced cumulative regret.

We elaborate on the power of the proposed double sampling technique by evaluating the relative difference between the averaged cumulative regrets of our algorithm and Thompson sampling, with respect to the Kullback-Leibler divergence. The minimum arm-divergence is a proxy metric for bandit complexity, and a key term on bandit lower-bound regrets [Lai and Robbins(1985)]. In a sense, this divergence is parameter agnostic, as many parameter combinations for a given model may map to the same Kullback-Leibler divergence.



(a) Averaged cumulative regret with $A = 2$, $w_0 = (0.4\ 0.4)^\top$, $w_1 = (0.8\ 0.8)^\top$, $\sigma_0 = \sigma_1 = 0.2$.

(b) Averaged cumulative regret with $A = 2$, $w_0 = (0.5\ 0.5)^\top$, $w_1 = (0.8\ 0.8)^\top$, $\sigma_0 = \sigma_1 = 1$.

Figure 4: Averaged cumulative regret comparison (standard deviation shown as shaded region).

In particular, Fig. 5 contains average cumulative regret relative differences (as in Equation 32), over 5000 realizations of two-dimensional contextual linear Gaussian bandits with two arms, for per-dimension parameter w_i grids within $[-1, 1]$ with gaps of 0.1, and $\sigma \in [0.1, 1]$ with step size of 0.1.

We observe that double sampling provides significant regret reductions for all the parameterizations of the two-armed contextual linear Gaussian bandit. The regret improvement is most evident for models with significant Kullback-Leibler divergences, with cumulative regret reductions of up to 40%.

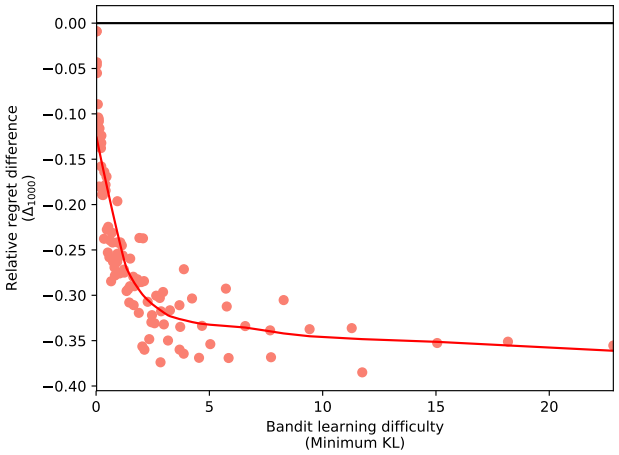


Figure 5: Relative average cumulative regret differences at $t = 1000$.

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 can provide improved regrets, which is in many application domains inexpensive in comparison with interacting with the world. Encouraged by these findings, we aim at extending this technique to other reward distributions and implementing it with datasets from domain applications in which actions are expensive relative to samples drawn from the current model.

References

[Agrawal and Goyal(2011)] Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. *CoRR*, abs/1111.1797, 2011. URL `http://arxiv.org/abs/1111.1797`.

[Agrawal and Goyal(2012a)] Shipra Agrawal and Navin Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. *CoRR*, abs/1209.3352, 2012a. URL `http://arxiv.org/abs/1209.3352`.

[Agrawal and Goyal(2012b)] Shipra Agrawal and Navin Goyal. Further Optimal Regret Bounds for Thompson Sampling. *CoRR*, abs/1209.3353, 2012b. URL `http://arxiv.org/abs/1209.3353`.

[Auer et al.(2002)] Peter Auer, Nicolò Cesa-Bianchi, and Paul 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. URL `http://dx.doi.org/10.1023/A:1013689704352`.

[Bellman(1956)] Richard Bellman. A Problem in the Sequential Design of Experiments. *Sankhya: The Indian Journal of Statistics (1933 - 1960)*, 16(3/4):221–229, 1956. URL `http://www.jstor.org/stable/25048278`.

[Bernardo and Smith(2009)] José M. Bernardo and Adrian F.M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, 2009. ISBN 9780470317716. doi: 10.1002/9780470316870. URL `http://onlinelibrary.wiley.com/book/10.1002/9780470316870`.

[Brezzi and Lai(2000)] Monica Brezzi and Tze Leung Lai. Incomplete Learning from Endogenous Data in Dynamic Allocation. *Econometrica*, 68(6):1511–1516, 2000. ISSN 1468 0262. doi: 10.1111/1468-0262.00170. URL `http://dx.doi.org/10.1111/1468-0262.00170`.

[Brezzi and Lai(2002)] Monica Brezzi and Tze Leung 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. URL `http://www.sciencedirect.com/science/article/pii/S0165188901000288`.

[Chapelle and Li(2011)] Olivier Chapelle and Lihong 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. URL http://papers.nips.cc/paper/4321-an-empirical-evaluation-of-thompson-sampling.pdf.

[Gittins(1979)] 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. URL http://www.jstor.org/stable/2985029.

[Korda et al.(2013)] Nathaniel Korda, Emilie Kaufmann, and Rémi 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. URL http://papers.nips.cc/paper/5110-thompson-sampling-for-1-dimensional-exponential-family-bandits.pdf.

[Lai(1987)] Tze Leung Lai. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. *The Annals of Statistics*, 15(3):1091–1114, 1987. ISSN 00905364. URL http://www.jstor.org/stable/2241818.

[Lai and Robbins(1985)] Tze Leung Lai and Herbert 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. URL http://dx.doi.org/10.1016/0196-8858(85)90002-8.

[Li et al.(2010)] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. *CoRR*, abs/1003.0146, 2010. URL http://arxiv.org/abs/1003.0146.

[Maillard et al.(2011)] Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences. In *Conference On Learning Theory*, 2011. URL http://hal.inria.fr/inria-00574987/fr/.

[Robbins(1952)] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, (58):527–535, 1952. doi: https://doi.org/10.1090/S0002-9904-1952-09620-8.

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

[Russo and Roy(2014)] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014. doi: http://pubsonline.informs.org/doi/abs/10.1287/moor.2014.0650.

[Russo and Roy(2016)] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016. URL http://www.jmlr.org/papers/volume17/14-087/14-087.pdf.

[Scott(2010)] Steven 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. URL http://dx.doi.org/10.1002/asmb.874.

[Scott(2015)] Steven L. Scott. Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, 31:37–49, 2015. URL http://onlinelibrary.wiley.com/doi/10.1002/asmb.2104/abstract. Special issue on actual impact and future perspectives on stochastic modelling in business and industry.

[Sutton and Barto(1998)] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press: Cambridge, MA, 1998. URL https://mitpress.mit.edu/books/reinforcement-learning.

[Thompson(1933)] 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. ISSN 00063444. URL http://www.jstor.org/stable/2332286.

[Thompson(1935)] William R. Thompson. On the Theory of Apportionment. *American Journal of Mathematics*, 57(2):450–456, 1935. ISSN 00029327, 10806377. URL http://www.jstor.org/stable/2371219.