Piece-wise stationary multi-armed bandits

Final project

Marc Agustí (marc.agusti@barcelonagse.eu)
Patrick Altmeyer (patrick.altmeyer@barcelonagse.eu)
Ignacio Vidal-Quadras Costa (ignacio.vidalquadrascosta@barcelonagse.eu)

26 June, 2021

1 Introduction

The multi-armed bandit (MAB) is a problem in reinforcement learning that focuses on how to solve the exploration-exploitation dilemma (Sutton and Barto 2018). Each of the arms has a probability of succeeding which is modelled by a Bernoulli distribution with a parameter p. Most of the theory around multi-armed bandits covered in class and its respective implementations assume stationary on the arms, that is, the probability of an arm succeeding does not change through time. However, in most real life settings, this strong assumption is not satisfied (Raj and Kalyani 2017).

For instance, consider the problem deciding which news to put in the front page of a news paper that will capture the attention of as many readers as possible. In order to model the response of the reader to the news shown, one can use a Bernoulli distribution where the p describes the probability that the user clicks on the news link. In the stationary setup, this probability is assumed to be constant, which is unrealistic: there are trends that lead to some articles being more popular during some period and less popular during other times. For instance, during the Eurocup, an article on football can be predicted to have a lot of clicks, however once the Eurocup is over and friendly games take over, an article on football might not be as interesting anymore and thus getting fewer clicks.

To this end, in this project we explore different strategies that have been proposed and tested in order to deal with the complication of non-stationarity. We compare the different strategies empirically. The remainder of this note is structured as follows: in section 2 we briefly summarise a set of recent papers that have emerged from this line of literature. This will provide us with a set of difference strategies for solving non-stationary multi-armed bandits and serve as the foundation for an empirical investigation of their performance in section 3. Finally, in section 4 we discuss the empirical results and conclude.

2 Strategies for solving non-stationary MABs

Before diving into non-stationarity we briefly review a few of the conventional approaches that are used to solve stationary multi-armed bandits. Consider a multi-armed bandit with K all of which yield rewards that are generated by a Bernoulli distribution. Let $\mathbf{p}$ denote the vector of constant (stationary) Bernoulli probabilities where element p_k, $k = 1, ..., K$ indicates the unconditional probability that arm a_k yields a reward. Without loss of generality we will assume that rewards in period t are binary: $\mathbf{r}_t = [0, 1]^K$. When thinking about solving this stationary MAB, the goal is acquire good action value estimates $\hat{q}(a)$ for all $a_1, ..., a_K$ and choose the optimal one.

In this stationary context, the idea of the MAB is quite simple: as the bandit repeatedly chooses actions and observes rewards it learns about the different action values, as long as it explores enough. In our case the true action value simply corresponds to its Bernoulli probablitiy. Figure 1 shows a sketch of the learned posteriour distributions attached to three different arms with true action values 0.2, 0.5, and 0.8. The general idea underlying

all strategies that aim to solve stationary MABs is that over time the bandit's uncertainty around the arms' action values decreases, provided it explores enough. Eventually then the bandit learns which is the optimal arm. Good strategies for solving stationary MABs adequately balance the trade-off between **exploring** and hence acquiring good action value estimates and **exploiting** arms with high estimated action values.

Figure 1: Sketch of estimated action value distributions for stationary probabilites.

Two of the simplest approaches to stationary MABs are referred to as ϵ-first and ϵ-greedy (Sutton and Barto 2018). The former is intertemporal composition of a random and a deterministic choice: it imposes that up until a certain point in the sampling period arms are chosen at random and thereafte the arm with the highest estimated action value is chosen. In particular, let $\tilde{a}_t$ denote a random draw from $\{a_1, ..., a_K\}$ in time period t and let T denote the total number of time periods. Then the ϵ-first approach can be formally defined as follows:

$$a_t = \begin{cases} \tilde{a}_t & \text{if } \frac{t}{T} < \epsilon \\ \arg\max_a \hat{q}_t(a) & \text{otherwise.} \end{cases} \tag{1}$$

The plain-vanilla ϵ-greedy approach is similar in that it is also a composition of a random and a deterministic choice. Formally we have

$$a_t = \tilde{a}_t I_t + (1 - I_t) \arg\max_a \hat{q}_t(a) \tag{2}$$

where $I_t \sim \text{Bern}(\epsilon)$ determines in each period whether the random a deterministic choice is applied. Note that in contrast to ϵ-first, the rule defined in (2) allows from exploration throughout the sampling period.

Another popular and simple choice for solving stationary MABs is **Softmax**.

$$P(a_t = a) = \frac{\exp(\frac{\hat{q}_t(a)}{\tau})}{\sum_{k=1}^{K} \frac{\hat{q}_t(a_k)}{\tau})} \tag{3}$$

The idea of softmax is very intuitive: the arm is chosen each period according to a probably distribution. This probabilty distribution gives more weight to arms that have been performing well. In this case we also allow for exploration throughout the sample.

All of these simple approaches can perform very well in the context of stationary MABs, but cumulative regret is typically a linear function of time. Algorithms that achieve sub-linear regret work under the premise of **optimism in the fact of uncertainty**: they exploit uncertainty in that they typically favour choosing arms that the bandit is uncertain about. The most common approaches are **Upper Confidence Bounds** (UCB) and **Thompson Sampling**. They largely differ in their approach towards quantifying uncertainty.

The general UCB rule for choosing arms can be defined as

$$a_t = \arg\max_a \left(\hat{q}_t(a) + \hat{U}_t(a) \right) \tag{4}$$

where the exact specification of the upper confidence bound $\hat{U}(\cdot)$ is at the researcher's discretion. Typical choices for the functional form of $\hat{U}(\cdot)$ decrease in t and and the number of times m_t that any given arm a has already been observed, for example $\hat{U}_t(\hat{a}) = \sqrt{\frac{2 \log t}{m_t}}$ (Auer, Cesa-Bianchi, and Fischer 2002). Another possible choice derived from Chernoff's bound and used in Chapelle and Li (2011) is $\hat{U}_t(\hat{a}) = \frac{k_t}{m_t} + \sqrt{\frac{2 \frac{k_t}{m_t} \log \frac{1}{\delta}}{m_t}} + \frac{2 \log \frac{1}{\delta}}{m_t}$ where $\delta = \sqrt{\frac{1}{t}}$ and k_t in our context denotes the number of times that any given arm has yielded a reward up until time period t.

Thompson Sampling instead takes a Bayesian approach towards quantifying uncertainty and has been shown to outperform UCB in the context of stationary multi-armed bandits (Chapelle and Li 2011). Formally we have for that

$$a_t = \arg\max_a f_{\mathbf{w}}(a), \quad \mathbf{w} \sim P(\mathbf{w}|\alpha + n_{a,1}, \beta + n_{a,0}) \tag{5}$$

where $P(\mathbf{w}|\alpha + n_{a,1}, \beta + n_{a,0}) = \prod_{k=1}^{K} \text{Beta}(w_k|\alpha + n_{a,1}, \beta + n_{a,0})$, $n_{a,1}$ here denotes the number of times arm a has yielded a reward and $n_{a,0}$ denotes the number of times a failed to yield a reward. The term $f_{\mathbf{w}}(a)$ just corresponds to the posteriour mean and hence the bandit's belief about the action value of arm a.

All of these strategies can in principal be used to solve the stationary Bernoulli MAB with the latter typically performing much better than the simple approaches we initially introduced. But it should by now be clear that the assumption of stationarity is crucial for these strategies to work well. Suppose now that instead of each arm having stationary reward probabilities these parameters change over time. In that case we would like to ensure that the bandit always has the capacity to update their beliefs about the action values. In particular, in a non-stationary or piece-wise stationary environment the bandit should be able to pick up on structural changes in the underyling probability distributions.

Figure 2 presents a sketch of action value distributions in a piece-wise stationary environment. Here we have imposted that in period 2 the true action values change to 0.95, 0.7, and 0.3

from their initial values. In period 3 and 4 they change to 0.05, 0.65, and 0.35 and 0.5, 0.6, and 0.95, respectively. The plotted distributions in Figure 2 can best be thought of as describing posteriour beliefs of a bandit that has learned these distributions independently during the different time periods. Of course, this is the ideal scenario, but not realistic: in practice the bandit enters each new period with prior beliefs shaped during the previous period. For example, a bandit that has adequately explored during period 1 enters period 2 thinking that arm 3 is the optimal choice. At the beginning of period 2 that is no longer true: in fact, arm 3 is now a very poor choice. But this news is only gradually revealed to the bandit as it observes new rewards (or lack thereof if it continues to choose arm 3). A good bandit therefore needs to be capable to quickly unlearn their prior beliefs in light of the new information.

Strategies designed to deal with non-stationary MABs incorporate these ideas. In particular, these strategies are mainly based on modifications of the aforementioned algorithms. The following two subsections discuss two interesting approaches and present the underyling methodology.

2.1 Discounted Thompson Sampling

We have based the evaluation of the Discounted Thompson Sampling (DTS) on Raj and Kalyani (2017). The key idea of DTS is to systematically increase the variance of the prior distributions maintained for unexplored arms. Formally, we have that

$$a_t = \arg\max_a f_{\mathbf{w}}(a), \quad \mathbf{w} \sim P(\mathbf{w}|\alpha + n_{a,1}, \beta + n_{a,0}) \tag{6}$$

where $P(\mathbf{w}|\alpha + n_{a,1}, \beta + n_{a,0}) = \prod_{k=1}^{K} \text{Beta}(w_k|\alpha + n_{a,1}, \beta + n_{a,0})$

like for standard Thomspon Sampling. The main difference lies in the updating rule for the number of successes and the number of failures. Now, for each iteration we will update these two values in the following way:

$$n_{a^*,1} \leftarrow \gamma n_{a^*,1} + r_t$$

$$n_{a^*,0} \leftarrow \gamma n_{a^*,0} + (1 - r_t)$$

where a^* is the action chosen in iteration t, $\gamma \in (0, 1)$ is a discount factor, and

$$n_{a,1} \leftarrow \gamma n_{a,1}$$

$$n_{a,0} \leftarrow \gamma n_{a,0}$$

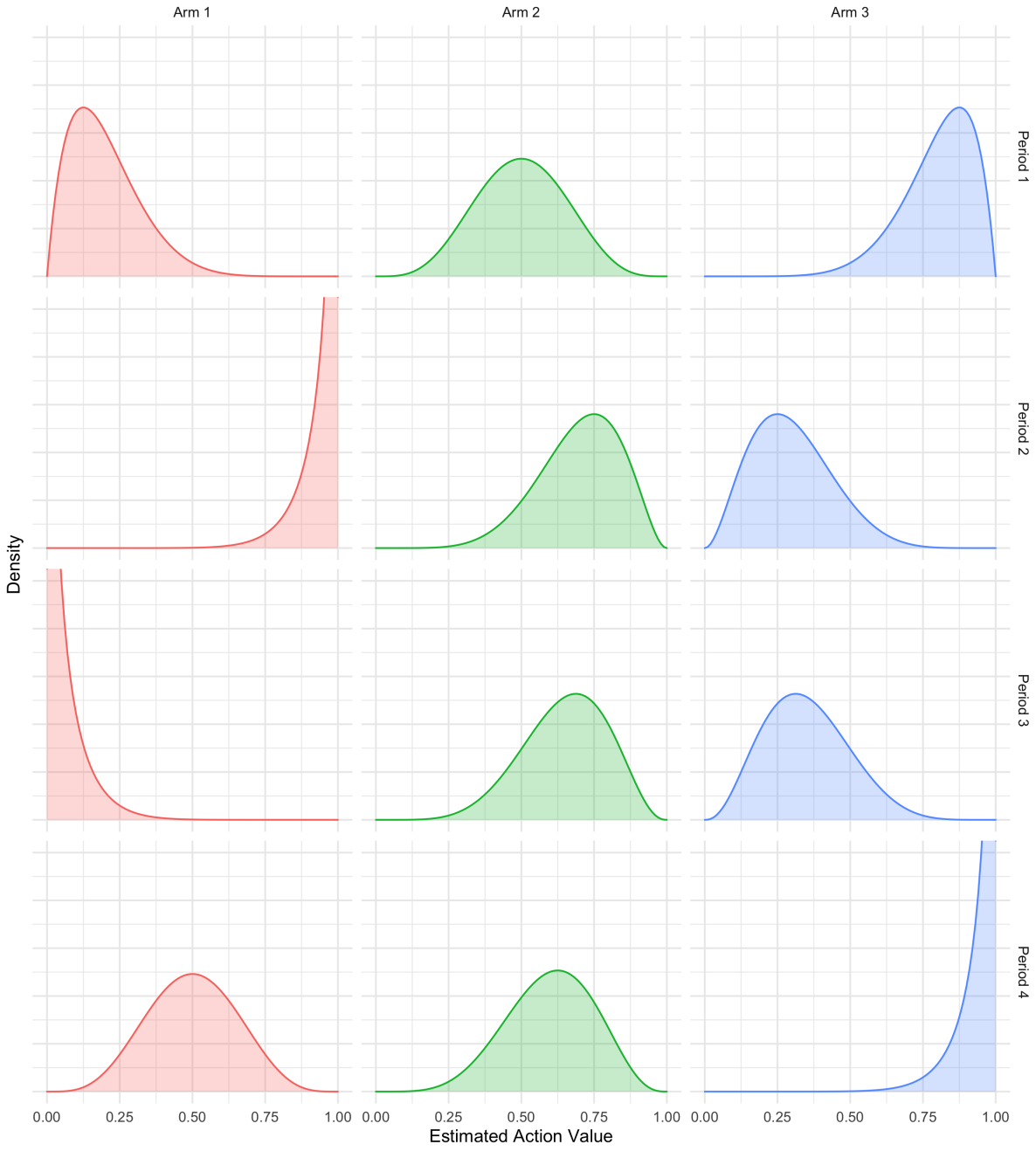

Figure 2: Sketch of estimated action value distributions for different periods. Bernoulli probabilities are assumed to be piece-wise stationary.

for all other actions. Therefore, as we can see, when sampling the posterior, we know incorporate a discount factor, which implies that observations far away in time have a lower weight than the most recent observations, so DTS works by discounting the effect of past observations. The algorithm updates parameters of all posterior distributions at every timestep. By increasing the variance of all arms, the probability of picking past inferior arms for exploration increases. However, by keeping the mean almost constant, the algorithm will not pick inferior arms too often.

2.2 Discounted Upper Confidence Bound

As stressed in Hartland et al. (2006) empirical evidence shows that UCB's exploration vs exploitation trade off is not appropriate for non-stationary for abruptly changing environments. To address this problem, Garivier and Moulines (2008) proposes the discounted UCB (DUCB). In order to estimate the expected reward, the DUCB policy averages past rewards with a discount factor giving more weight to recent observations, in line with the approach followed by DTS. This policy constructs an UCB $\bar{X}_t(\gamma, i) + c_t(\gamma, i)$ for the expected reward, where the discounted average is given by

$$\bar{X}_t(\gamma, i) = \frac{1}{N_t(\gamma, i)} \sum_{s=1}^{t} \gamma^{t-s} X_s(i) \mathbb{I}_{a_s=i}$$

where $X_s(i)$ is the reward obtained by arm i at timestep s, and where

$$N_t(\gamma, i) = \sum_{s=1}^{t} \gamma^{t-s} \mathbb{I}_{a_s=i}$$

and the discounted padding function is defined as

$$c_t(\gamma, i) = 2B\sqrt{\frac{\xi log(n_t)}{N_t(\gamma, i)}}$$

and $n_t = \sum_{i=1}^{K} N_t(\gamma, i)$. First of all, note that for $\gamma = 1$, you get standard UCB. However, the main contribution of Garivier and Moulines (2008) is the Sliding-Window UCB. They propose a methodology where instead of averaging the rewards over all past with a discount factor they use a local empirical average of the observed rewards, by using only the τ last plays. Specifically, this algorithm constructs an UCB $\bar{X}_t(\tau, i) + c_t(\tau, i)$ where the local empirical average is given by

$$\bar{X}_t(\tau, i) = \frac{1}{N_t(\tau, i)} \sum_{s=t-\tau+1}^{t} \gamma^{t-s} X_s(i) \mathbb{I}_{a_s=i}$$

and the padding function is defined as

$$c_t(\tau, i) = 2B\sqrt{\frac{\xi log(min(t, \tau))}{N_t(\gamma, i)}}$$

3 Empirical investigation

We now move on to comparing the different proposed methodologies. As the testing ground we will use the piece-wise stationary environment that we already introduced in the previous section.

In Figure 3 we can observe the performance of the different sampling methods described in the previous sections. The first thing we notice is that the sampling methods that incorporate a discounting factor yield the best performance in the non-stationary setting as expected. As noted above, the introduction of a discount factor implements the premise that realizations that lie far in the past should be weighted less when updated beliefs about action values. Or in other words, recent information has more relevance when estimating the expected reward of any given arm. Both Discounted UCB and Discounted Thompson appear to quickly pick up on the change in the underlying Bernoulli distribution and therefore acummuluate less regret during these sample periods than conventional UCB and Thompson. The Sliding Window approach is very similar to the discounted UCB approach, which is reflected in its similar performance.

When looking at the conventional UCB more closely we can observe that it is overall the best sampling method during the first half of the sample period. When the first structural change happens, it reacts relatively swiftly but initially incurs a steep increase in cumulative regret. The same is observed for Thompson Sampling, though it is slower to react to the changes and therefore performs worse. In the third period we see something interesting: we can see that UCB is much slower to react to the probability changes. This is likely due to the fact that in this periods all the arms have a similar probability of reward. It therefore takes longer to gain certainty about the action values. Consistent with that logic, the final change in probabilities clearly favours one particular arm and as consequence even the conventional sampling methods accumulate only little regret.

As mentioned in previous sections, conventional Thompson sampling tends to perform better than conventional UCB in a stationary setting, but as evident from the results here this is no longer the case in the non-stationary setting. If past beliefs are not discounted, then the Bayesian way for sampling arms does not appear to provide the necessary flexibility to unlearn prior beliefs. On the other hand, our implementation of the UCB adapts much better to changes in the distribution and therefore achieves better cumulative regret in the non-stationary setting.

Finally, softmax, ϵ-first and ϵ-greedy do very poorly as one would have expected. In the case of softmax, we can see that choosing the arm with respect to probabilities that are proportional to the mean rewards also does not achieves a good performance. The simple ϵ-first only explores during the pure exploration phase and is therefore completely insensitive

to structural changes that occur during the pure exploitation period. Performance for ϵ-greedy is marginally better, likely due to the fact that it randomly explores throughout the entire sample period.

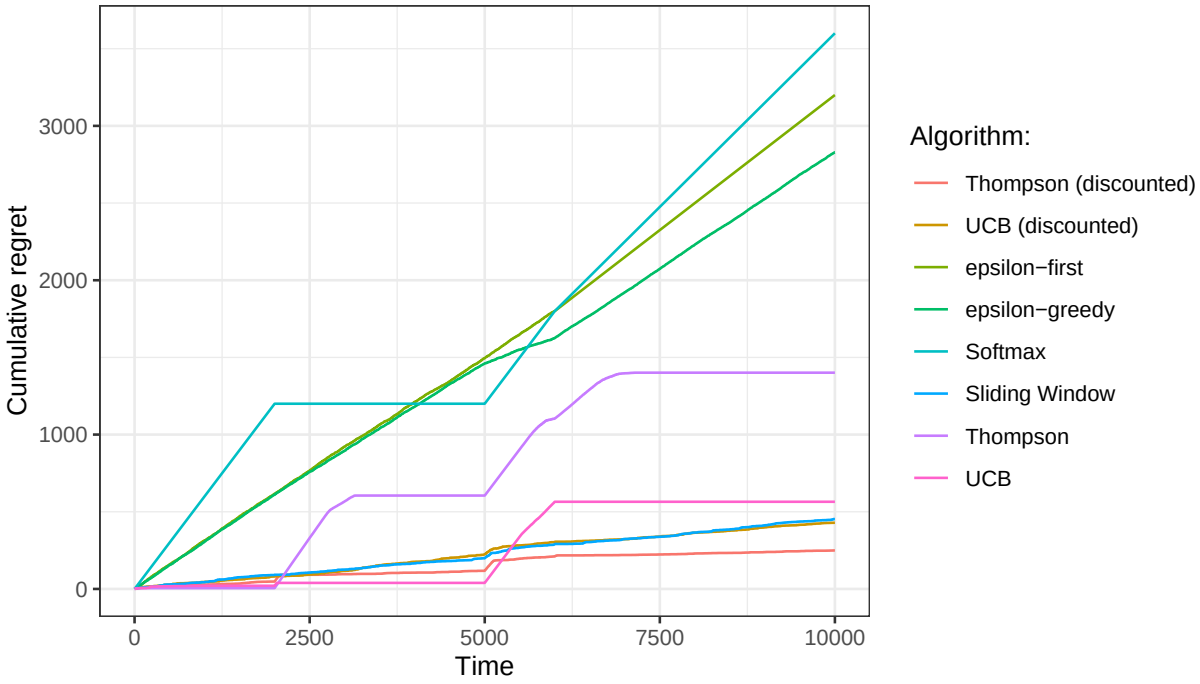

Figure 3: Benchmarking strategies for solving a piece-wise stationary Bernoulli multi-armed bandit.

3.1 Discounted Thompson Sampling

Now that we have seen some preliminary evidence in favour of discounted sampling methods for non-stationary multi-armed bandits, in this section we will investigate the empirical performance of discounted Thompson Sampling (DTS) a little further. To this end we have implemented the algorithm in `C++` for greater computational efficiency.

As a first exercise, we repeatedly apply DTS for subperiods of the sample and inspect the posteriour distributions for the action values. In particular, we train the TDS bandit independently on the first period only ($T_1 = 2000$), then until the end of the second period ($T_2 = 5000$), then until the end of the third period ($T_3 = 6000$) and finally over the entire sample ($T_4 = 10000$). We do this 100 times and each time return the discounted number of successes and failures. Average discounted counts across all 100 simulations then serve as our input for the Beta distribution, from which we draw repeatedly in order to obtain an empirical distribution for the posterior.

Since DTS has been shown to perform well, we expect that the shapes of the resulting distributions should be broadly consistent with the distributions we sketched above in Figure 2. The results are shown in Figure 4 in the appendix. Indeed, they look very similar to the

sketched distributions which were based on the true underlying Bernoulli probabilities. In other words, the DTS bandit appears to be capable of forming posteriour beliefs that are very much in line with reality, even though after each structural break it first has to unlearn its prior beliefs.

Finally, recall that the discount factor γ involved in DTS is a free parameter. To the best of our knowledge, no guiding principles have been proposed in order to choose optimal values, although this would be an interesting theoretical question. Instead, the discount factor needs to be optimized through cross-validation. To this end, Figure 5 in the appendix shows how the cumulative regret varies with the discount factor. Upon visual inspection it appears that a value close to 0.997 is optimal in this particular context. A choice of 1 corresponds to the non-discounted Thompson Sampler and yield poor results. But performance also tends to decrease for choices lower than 0.997, demonstrating the need for tuning.

4 Conclusion

In this short note we have explored how non-stationarity affects the sampling methods seen in class and other popular methods in the literature to choose actions in the context of MABs. We have explored the performance of ϵ-first, ϵ-greedy, softmax, UCB, Sliding Window UCB, Discounted UCB, Thompson Sampling and Discounted Thompson Sampling.

We provide empirical evidence demonstrating that methods that allow to give more relevance to present samples as opposed to samples from far in the past by including a discount factor perform very well in the non-sationary setting. On the other hand, those methods that fail to take non-stationarity into account unsurprisingly yield poor results. Among all sampling methods tested here, the one that achieves the best performance is Discounted Thompson Sampling (DTS). We have illustrated that DTS is capable of learning adequate posterior beliefs about action values even as it involves unlearning prior beliefs. Finally, we have also shown that the choice of the discount factor matters and practitioners should therefore carefully tune it.

References

Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. "Finite-Time Analysis of the Multiarmed Bandit Problem." *Machine Learning* 47 (2): 235–56.

Chapelle, Olivier, and Lihong Li. 2011. "An Empirical Evaluation of Thompson Sampling." *Advances in Neural Information Processing Systems* 24: 2249–57.

Garivier, Aurélien, and Eric Moulines. 2008. "On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems." *arXiv Preprint arXiv:0805.3415*.

Hartland, Cédric, Sylvain Gelly, Nicolas Baskiotis, Olivier Teytaud, and Michele Sebag. 2006. "Multi-Armed Bandit, Dynamic Environments and Meta-Bandits."

Raj, Vishnu, and Sheetal Kalyani. 2017. "Taming Non-Stationary Bandits: A Bayesian Approach." *arXiv Preprint arXiv:1707.09727*.

Sutton, Richard S, and Andrew G Barto. 2018. *Reinforcement Learning: An Introduction*. MIT press.

A Additional Figures

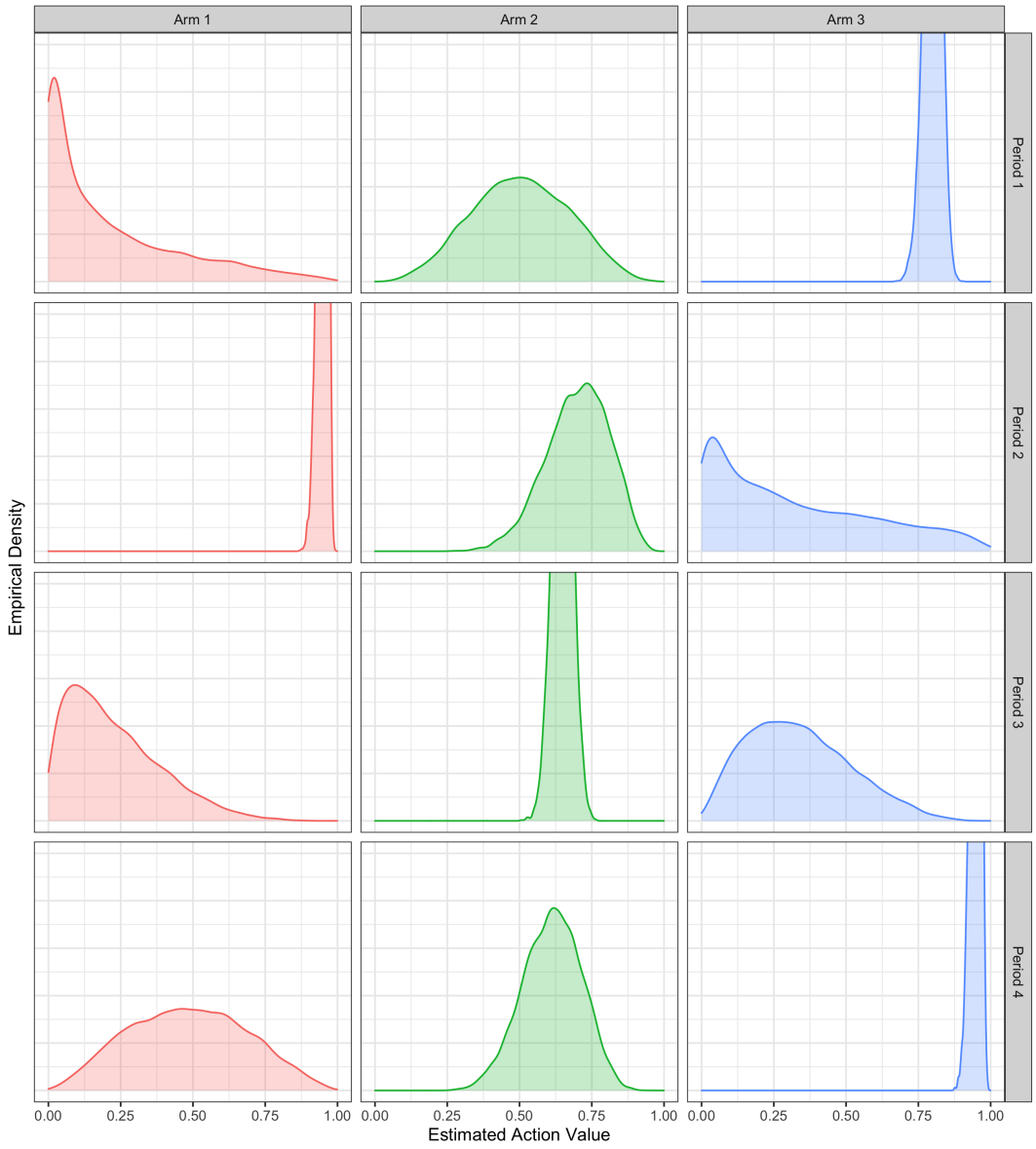

Figure 4: Empirical distribution of estimated action values for different periods and corresponding piece-wise stationary probabilites.

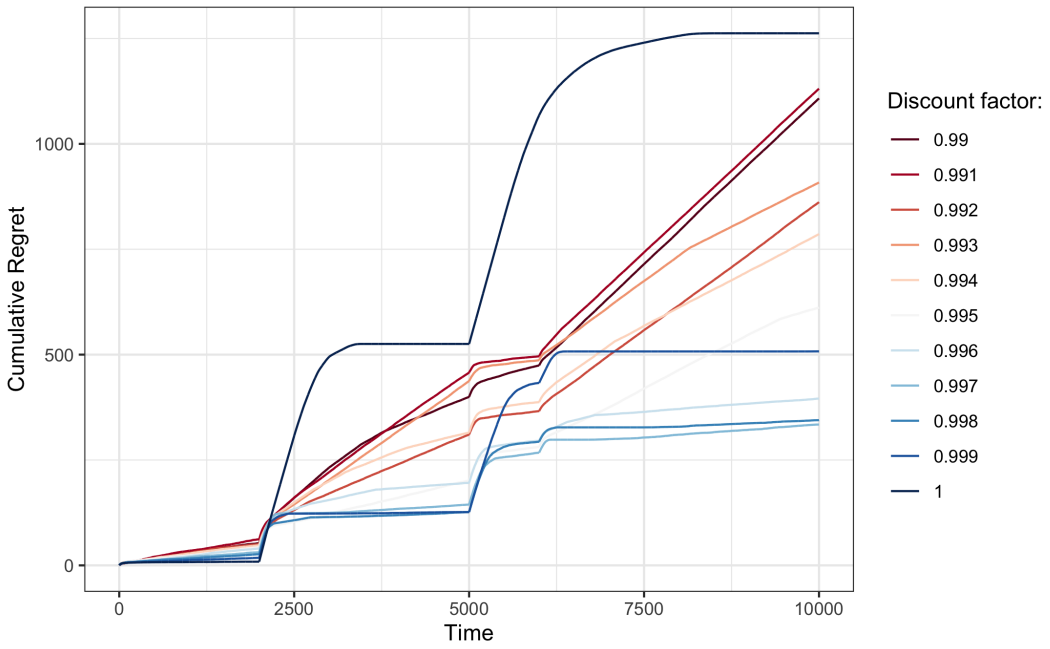

Figure 5: The effect of the discount factor on cumulative regret.