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. ...

NOTE: All to complete

Raj and Kalyani (2017) - **Taming Non-stationary Bandits: A Bayesian Approach** Besbes, Gur, and Zeevi (2014) - **Stochastic multi-armed-bandit problem with non-stationary rewards** Gupta, Granmo, and Agrawala (2011) - **Thompson Sampling for Dynamic Multi-armed Bandits** Garivier and Moulines (2008) - **On upper-confidence bound policies for non-stationary bandit problems**

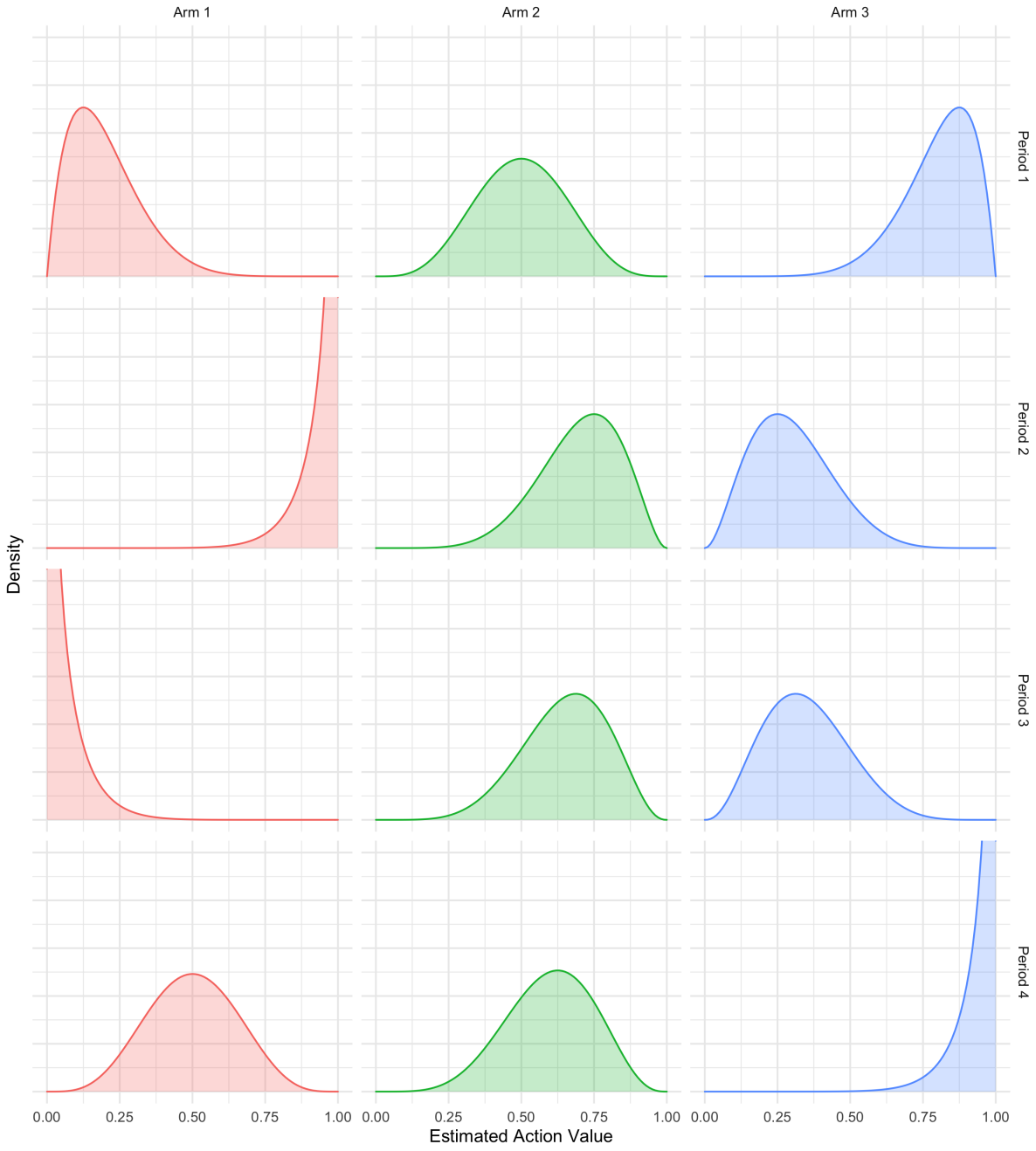

Figure 2: Sketch of estimated action value distributions for different periods. Bernoulli probabilities are assumed to be piece-wise stationary.

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. For reference, Table 1 in the appendix summarises how the Bernoulli probability vary across the different time periods.

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 ammuluate less regret during these sample periods than conventional UCB and Thompson. When looking at UCB in particular we can observe that it is the best sampling method in the beginning of the sample. 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. Consisitent 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, Thompson sampling performs better than UCB in a stationaray setting, but as seen in the results it is no longer the case in the non-sataionary setting. The bayesian way for sampling arms does not provied the necessary felxibility to the method in order to correctly identidy and adapt its smapling behaviour to the change in the underlying expected rewards of the differnt arms. On the other hand, the simple and intuitive approch of UCB of considering the upper confiende bound given the current estimate of the variance (driven by the amount of times we have explored the arm) and its estimated expected reward, adapts much better to changes in the distribution and therefore achiveing much bettter cunmmulative regret in the non stationary setting.

Finally, both softmax and ϵ-first do very poorly as one might have expected. The latter only explores during the pure exploration phase and is therefore completely insensitive to structural changes that occur during the pure exploitation period. In the case of softmax, we can see that choosing the arm with respect to probilities that are proportional to the mean rewards also does not achieves a good performance.

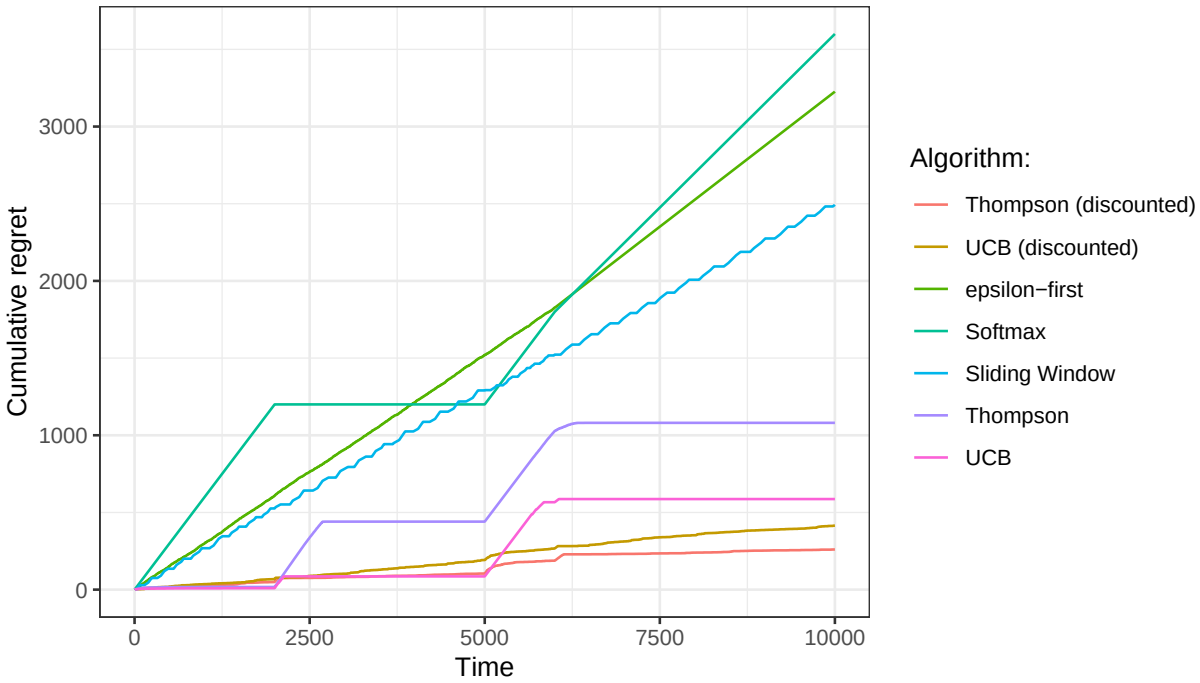

Figure 3: Benchmarking strategies for solving a piece-wise stationary Bernoulli multi-armed bandit.

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 a little further. To this end we have implemented the algorithm in `C++` for greater computational efficiency.

4 Discussion

NOTE: All to complete

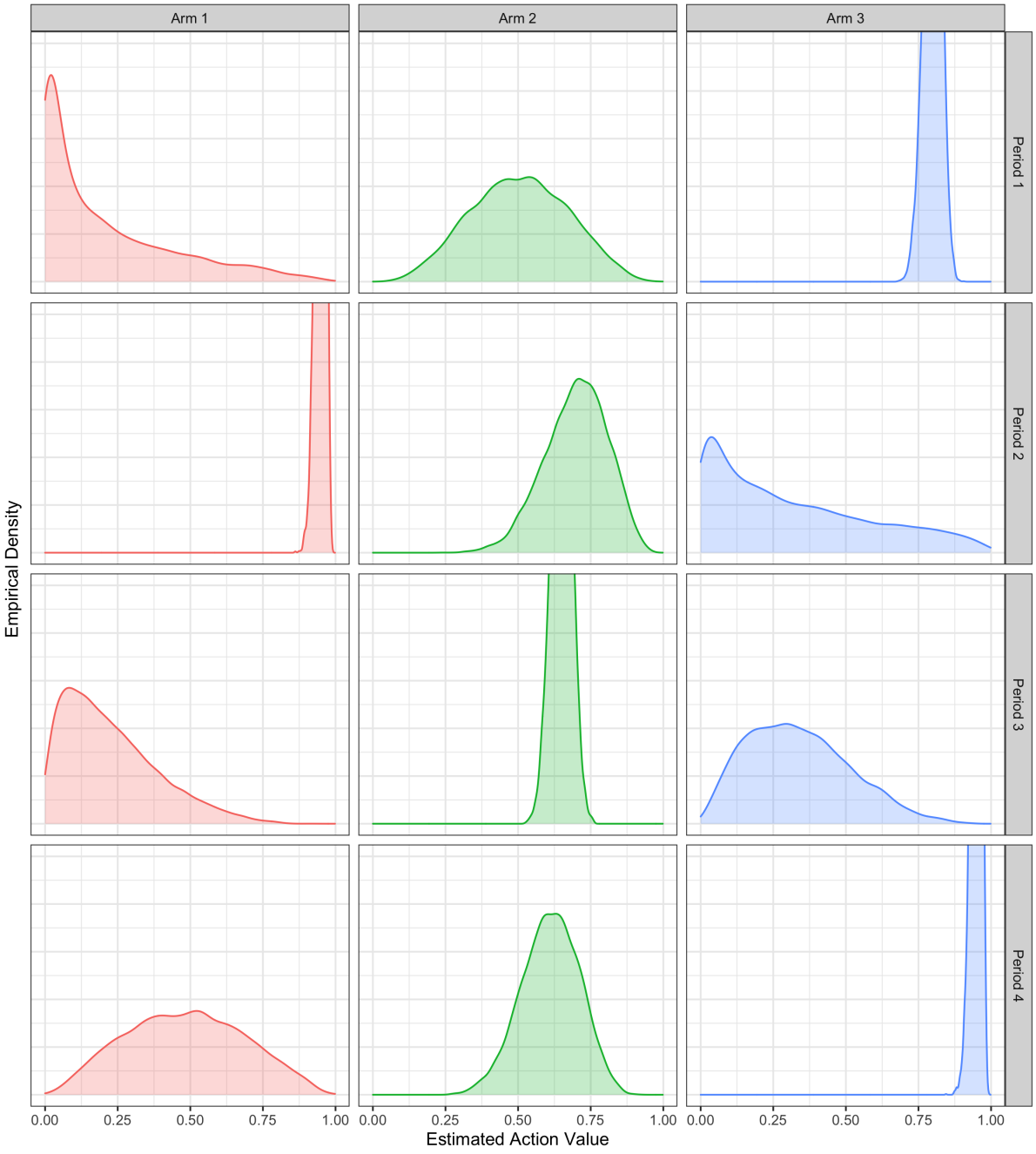

Figure 4: Empirical distribution of estimated action values for different periods and corresponding piece-wise stationary probabilites.

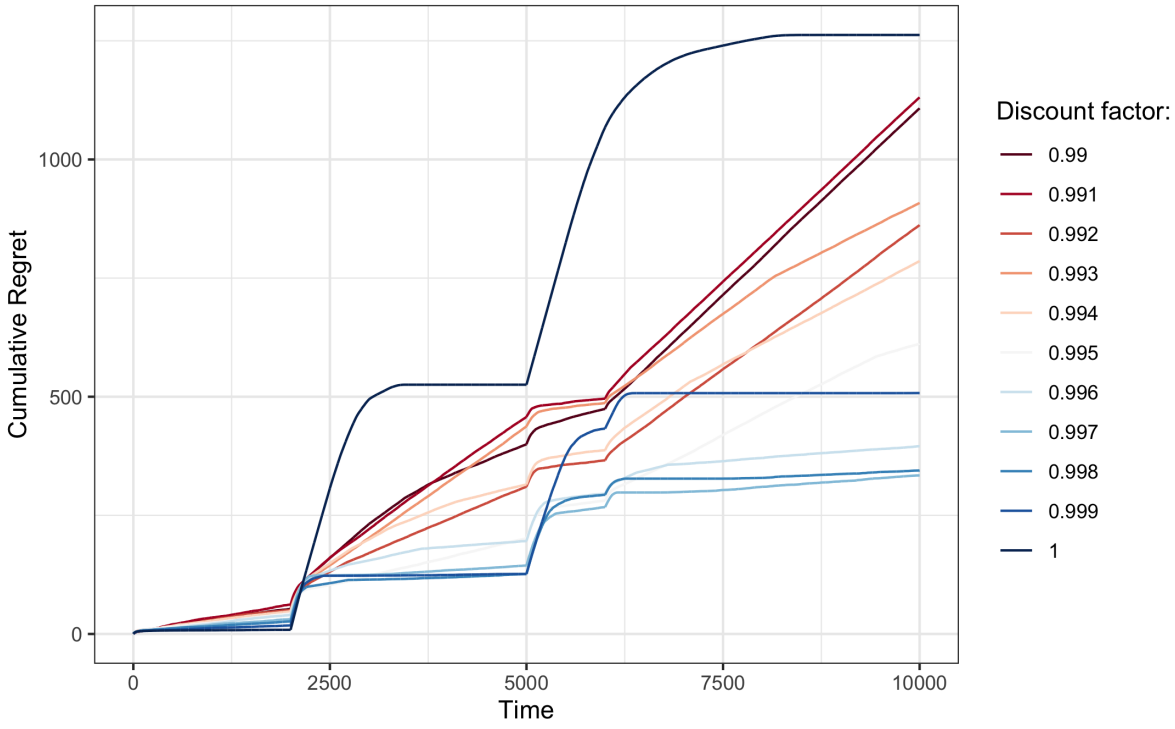

Figure 5: The effect of the discount factor on cumulative regret.

References

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

Besbes, Omar, Yonatan Gur, and Assaf Zeevi. 2014. "Stochastic Multi-Armed-Bandit Problem with Non-Stationary Rewards." *Advances in Neural Information Processing Systems* 27: 199–207.

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*.

Gupta, Neha, Ole-Christoffer Granmo, and Ashok Agrawala. 2011. "Thompson Sampling for Dynamic Multi-Armed Bandits." In *2011 10th International Conference on Machine Learning and Applications and Workshops*, 1:484–89. IEEE.

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*.

Table 1: A piece-wise stationary Bernoulli multi-armed bandit.

Period	T	Arm 1	Arm 2	Arm 3
1	2000	0.20	0.50	0.80
2	5000	0.95	0.70	0.30
3	6000	0.05	0.65	0.35
4	10000	0.50	0.60	0.95

MIT press.

5 (APPENDIX) Appendix