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)

24 June, 2021

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**

1 Introduction

The Multi-Armed Bandit problem 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

The **Upper Confidence Bound** (UCB) approach to solving the multi-armed bandit problem involves ... (Sutton and Barto 2018) **Thompson Sampling** has been shown to outperform UCB in the context of stationary multi-armed bandits (Chapelle and Li 2011).

3 Empirical investigation

4 Discussion

References

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*. MIT press.