International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 04 | Apr 2023 www.irjet.net p-ISSN: 2395-0072

NON-STATIONARY BANDIT CHANGE DETECTION-BASED THOMPSON SAMPLING ALGORITHM: A REVIEW

Md Arif[1], Mr. Nadeem Ahmad[2]

[1]*M.Tech, Electronic and Communication Engineering, GITM, Lucknow, India*
[2]*Assistant Professor Electronic and Communication Engineering, GITM, Lucknow, India*

---***---

Abstract - *Thompson Sampling (TS) is a popular algorithm used in multi-armed bandit problems. In the standard setting, it assumes that the rewards associated with each arm are stationary, which means that the reward distribution remains fixed throughout the experiment. However, in many real-world scenarios, the reward distribution can change over time, and this is known as a non-stationary bandit problem. In this case, the traditional TS algorithm may not perform well. To address this issue, several extensions of the standard TS algorithm have been proposed for non-stationary bandits. One such extension is the Bayesian Online Changepoint Detection (BOCD) algorithm. BOCD uses a Bayesian framework to model the changes in reward distribution and adjust the exploration and exploitation trade-off accordingly. The BOCD algorithm maintains a posterior distribution over the possible locations of the change points in the reward distribution. At each time step, it uses this posterior to compute the probability that a change point has occurred. If the probability of a change point is high, the algorithm explores more to adapt to the new reward distribution. Otherwise, it exploits more to maximize its expected reward.*

Another extension of the standard TS algorithm for non-stationary bandits is the Dynamic Thompson Sampling (DTS) algorithm. DTS uses a sliding window approach to detect changes in the reward distribution. The algorithm maintains a separate posterior distribution for each window and selects the arm with the highest expected reward based on the posterior distribution of the current window. In summary, Thompson Sampling is a powerful algorithm for the multi-armed bandit problem, and several extensions can be used to handle non-stationary bandits. These extensions allow the algorithm to adapt to changes in the reward distribution over time and continue to make optimal decisions.

Key Words: Sampling, Algorithm for Non-Stationary Bandits, Stationary Bandits, detection based.

1. INTRODUCTION

The multi-armed bandit (MAB) framework is a classic problem in decision theory and reinforcement learning. It involves an agent (or decision-maker) who must choose between a set of actions (often called "arms"), each of which has an unknown reward distribution. The goal of the agent is to maximize its cumulative reward over time while balancing the exploration of new actions with the exploitation of actions that have already been found to be rewarding.

One common example of the MAB problem is the slot machine or "one-armed bandit" problem. In this scenario, the agent must choose between pulling the levers of several slot machines, each of which has a different payout distribution. The agent must decide how many times to pull each lever to maximize its total payout.

There are many variations of the MAB problem, each with different assumptions and objectives. One common approach is the epsilon-greedy algorithm, which chooses the action with the highest estimated reward with probability 1-epsilon, and chooses a random action with probability epsilon. This balances the exploitation of known rewarding actions with the exploration of new actions.

Other popular algorithms for the MAB problem include UCB1, Thompson Sampling, and EXP3. These algorithms differ in their assumptions about the distribution of rewards, and their strategies for balancing exploration and exploitation.

The MAB problem has many applications in fields such as advertising, finance, and healthcare. For example, in online advertising, advertisers must decide which ads to display to maximize click-through rates while balancing the need to explore new ads with the need to exploit effective ads.

2. UPPER CONFIDENCE BOUND (UCB) ALGORITHM

The Upper Confidence Bound (UCB) algorithm is a popular algorithm for multi-armed bandit problems, which are a class of sequential decision-making problems. In the multi-armed bandit problem, a decision maker must repeatedly choose between a set of actions (or "arms") with uncertain rewards, to maximize their total reward over time.

The UCB algorithm works by balancing the exploration of different actions with their potential rewards. At each time step, the algorithm chooses the action with the highest upper confidence bound, which is a measure of the uncertainty of the estimated reward for that action.

The upper confidence bound is calculated as the sum of two terms: the estimated reward for the action and a confidence interval term that takes into account the uncertainty in the estimate. The confidence interval term typically grows over time to ensure that the algorithm continues to explore new actions, even as the estimates of the rewards for existing actions become more certain.

By balancing exploration and exploitation in this way, the UCB algorithm can achieve good performance in a wide range of multi-armed bandit problems.

3. ALGORITHM- DISCOUNTED THOMPSON SAMPLING (DTS)

Discounted Thompson Sampling (DTS) is a decision-making algorithm that is often used in the context of online advertising, recommendation systems, and other settings where there is uncertainty about the effectiveness of different actions or choices. DTS is a modification of the well-known Thompson Sampling algorithm, which is a Bayesian approach to decision-making that has been widely used in machine learning and artificial intelligence.

The basic idea behind DTS is to take into account the fact that the value of future rewards may decrease over time, due to factors such as discounting or decay. This means that it may be more important to take action now, rather than waiting for more information to become available. To incorporate this idea into the Thompson Sampling algorithm, DTS uses a discount factor that reduces the weight given to future rewards.

In DTS, each action or choice is associated with a probability distribution over the possible rewards. At each step of the algorithm, a random sample is drawn from each of these distributions, and the action with the highest sample is chosen. The distribution for each action is updated based on the observed reward, using Bayesian inference. The discount factor is applied to the observed reward before it is used to update the distribution.

One of the advantages of DTS is that it can handle non-stationary environments, where the underlying probabilities or reward distributions may change over time. The discount factor allows the algorithm to adjust to these changes, by placing more weight on recent observations.

DTS is effective in a variety of applications, including online advertising and recommendation systems. However, like all decision-making algorithms, its effectiveness depends on the specific context and the quality of the input data.

3.1.Stochastic Bandits

Stochastic bandits are a class of decision-making problems in which an agent must select actions from a set of options or arms, with uncertain rewards associated with each arm. In other words, the agent doesn't know the true reward distribution of each arm, but it can explore by taking different actions and observing the resulting rewards. The goal of the agent is to maximize the total reward it receives over a certain period, or to minimize the regret, which is the difference between the expected reward if the agent had chosen the optimal arm from the beginning and the actual reward obtained.

There are several algorithms to solve stochastic bandit problems, including the epsilon-greedy, UCB (Upper Confidence Bound), and Thompson sampling. Epsilon-greedy is a simple algorithm that selects the best arm with probability 1-epsilon and a random arm with probability epsilon. UCB is a more sophisticated algorithm that balances exploration and exploitation by selecting arms with the highest upper confidence bounds. Thompson sampling is a Bayesian algorithm that randomly samples reward distributions for each arm and selects the arm with the highest expected reward according to the samples.

Stochastic bandits have numerous applications in various fields, including recommender systems, online advertising, and clinical trials.

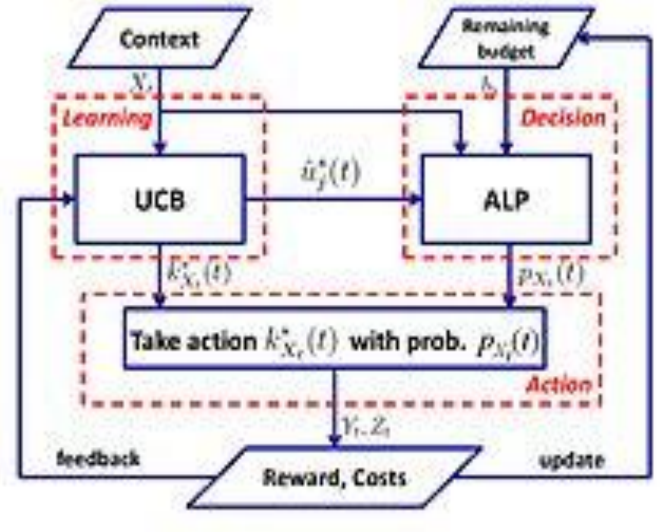

Figure-1: Stochastic Bandits

4. LITERATURE REVIEW

In the section of the literature review, we have studied the previous research paper related to Thompson sampling for different decision-making, the summary of all previous research papers is given below:

Joseph: Every genuinely autonomous system must have the capacity to make choices. One approach to finding such ways is via experience-based learning of decision-making strategies, since this kind of learning may help uncover strategies that are robust against the influence of noise and flexible enough to be used in a variety of settings. The multi-armed bandit problem and the best arm identification problem have been used as starting points for our exploration of this area (and related variations). After much trial and error, we've found that Thompson Sampling is an effective method for dealing with the multi-armed bandit algorithm. This solution is optimal for the Bernoulli armed bandit issue because it meets the lower restriction on anticipated cumulative regret. In many circumstances, the cost of computing is minimal. By using conjugate priors, MCMC, or variational methods, Thompson Sampling may be used to model a broad range of decision-making settings, including contextual bandit issues. It is adaptable to include cross-hand dependencies and is resistant to reward delays. To extend the evidence of Agrawal and Goyal to the optimistic variant of the approach, we make certain adjustments to optimism such that it also fulfills the lower limit for cumulative regret. As an additional finding, we find that the Normal Thompson Sampling approach (one that employs the Normal distribution as the conjugate prior) outperforms the more standard Thompson Sampling strategy (which employs a Beta prior) in terms of empirical performance.

Mellor, Jonathan: Many methods that combine Thompson Sampling with Change Point detection have been explored in this work. In their designated roles, we have shown their utility. Bandit scenarios using real-world data sets like the Yahoo! dataset and the Foreign Exchange further highlight their usefulness. They are shown to fail the PASCAL challenge when pitted against comparably complex but carefully calibrated competing algorithms. Our research, however, lends credence to a pair of methods that are analogous to Adapt-EvE but only keep tabs on changes to the "best" arm: Global-CTS2 and PA-CTS2. Due to its adaptability, the model holds out possibilities for further improvement via the introduction of additional assumptions. It's also worth noting that other reward distributions beyond Bernoulli's are completely OK. False alarms are a common issue with change point detection methods, but a Bayesian approach might help avoid them. Given that they are derived from fundamental models, the algorithms have theoretical support; nonetheless, further research is needed before a complete theoretical account of their performance can be provided.

Alami: We propose Global-STS-CF, an extension of the Switching Corrupt Bandit Problem using three distinct Thompson Sam plings. The experimental results demonstrate the superior performance of the suggested method. Notably, Global-STS-CF competes with an oracle, Corrupted Feedback, that is aware of the inflection points in advance, posing a challenge to Thompson sampling. These outcomes are a direct consequence of Global-STS-foundation CF's in the Bayesian idea of following the most knowledgeable individuals throughout the world, which enables us to detect and react to shifts in the landscape with remarkable efficiency. By keeping an expert distribution on a per-arm basis, the suggested approach may be easily adapted to the Per-arm Switching Corrupted Multi-Armed Bandit. Next, we'll examine the Global-STS-CF via the lens of pseudo-cumulative regret.

Gourab: To keep tabs on the dynamic two-armed bandit problem environment, we investigate a change-detection-based Thompson Sampling approach (TS-CD). We have established the minimum stationary regime time window for TS-CD, making it possible to detect the transitions as soon as they occur. Our results show that the proposed TS-CD algorithm converges to asymptotic optimality over a wide range of update rates. To test the efficacy of the strategy, we apply it to the RAT selection problem at the wireless network's edge. We have shown that TS-CD is superior to the standard max power band selection method and other bandit algorithms designed for unpredictable environments.

Maarten: We have investigated the OLTR issue in a dynamic setting where user tastes change quickly. Here, we provide cascading non-stationary bandits, an online-learning variation of the widely-used cascade model (CM) for predicting users' click behaviors. It has been suggested that the algorithms CascadeDUCB and CascadeSWUCB be used to solve it. They are proven to experience sub-linear remorse by our theoretical analysis. Our experimental results on the Yandex click dataset corroborate these theoretical predictions. Many new paths for the development of mobile OLTR are opened up. To begin with, we have just thought about the CM configuration. Future research should take into account other click models such as DBN [Chapelle and Zhang, 2009] that can process multiple clicks. The second area of interest was UCB-based policy. One alternative is to apply a policy from the softmax family [Besbes et al., 2014]. In this direction, it is possible to get upper limits that are insensitive to the number of transitions.

Giuseppe: We put forward a novel bandit method to solve the issue of learning in dynamic settings. The algorithm demonstrated superior performance over state-of-the-art solutions and the ability to adapt to various patterns of nonstationarity. All-Season is also far easier to use and more sturdy than its competitors, making it well-suited for use in industrial settings and requiring less upkeep. We speculate that there is more work to be done in addressing the model misspecification issue and choosing which models should be eliminated. In particular, we think that constructing the posterior prediction weights using the General Bayes technique (Knoblauch et al., 2019) might be a more robust option.

Gourab et.al: To account for the unpredictability of real-world environments, we provide a KS-test-based change detection technique in the context of the Multi-Agent-Based (MAB) architecture. We use our KS-test-inspired, actively adaptable TS algorithm, TS-KS, for the MAB problem. TS-KS has a sub-linear regret in the two-armed bandit problem. It's important to note that the proposed approach may detect a change even when more sophisticated methods based on mean estimates fail. As seen in two instances, we demonstrate that the TS-KS algorithm outperforms both the actively adaptive TS-CD technique and the passively adaptive D-TS strategy. To add, the results of the portfolio optimization case study demonstrate that TS-KS is competitive with other leading forecasting algorithms, like Facebook-PROPHET and ARIMA.

Zhang: Due to the inherent uncertainty of real-world settings, we provide a KS-test-based change detection approach within the framework of the Multi-Agent-Based (MAB) architecture. As an application of our KS-test-inspired, actively adaptive TS algorithm, TS-KS, we tackle the MAB issue. In the two-armed bandit dilemma, TS-KS experiences a sub-linear regret. Note that the suggested technique may succeed when more advanced approaches based on mean estimation fail to spot a shift. Using two examples, we show that the TS-KS algorithm outperforms the active TS-CD method and the passive D-TS approach. Furthermore, the portfolio optimization case study shows that TS-KS can hold its own against other top forecasting algorithms like Facebook-PROPHET and ARIMA.

5. CONCLUSION

After studying the above literature review, we found some conclusion, that is given here. Thompson sampling, also known as the Bayesian bandit algorithm, is a popular algorithm used for decision-making problems in various fields such as marketing, medicine, and engineering. It is a probabilistic algorithm that balances exploration and exploitation to find the optimal solution for a given problem. Thompson sampling can be used to optimize decision-making in robotics applications such as autonomous vehicles and drones. Thompson sampling can be used to optimize the recommendation process in recommender systems by choosing the most relevant items to recommend to users. Thompson sampling can be used to optimize the design of clinical trials by efficiently allocating patients to different treatment groups.

REFERENCE

[1] H. Mohanty, A. U. Rahman, and G. Ghatak. (2020) Thompson Sampling GoF. [Online]. Available: https://github.com/hardhik-99/ Thompsom Sampling GoF/

[2] D. Lee, N. He, P. Kamalaruban, and V. Cevher, “Optimization for reinforcement learning: From a single agent to cooperative agents,” IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 123–135, 2020.

[3] D. Bouneffouf and I. Rish, “A survey on practical applications of multi-armed and contextual bandits,” arXiv preprint arXiv:1904.10040, 2019.

[4] P. Englert and M. Toussaint, “Combined optimization and reinforcement learning for manipulation skills.” in Robotics: Science and systems, vol. 2016, 2016.

[5] G. Burtini, J. Loeppky, and R. Lawrence, “A survey of online exper⍰iment design with the stochastic multi-armed bandit,” arXiv preprint arXiv:1510.00757, 2015.

[6] A. Lesage-Landry and J. A. Taylor, “The multi-armed bandit with stochastic plays,” IEEE Transactions on Automatic Control, vol. 63, no. 7, pp. 2280–2286, 2017.

[7] S. Li et al., “Collaborative filtering bandits,” in Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, 2016, pp. 539–548.

[8] S. Buccapatnam, F. Liu, A. Eryilmaz, and N. B. Shroff, “Reward maxi⍰mization under uncertainty: Leveraging side-observations on networks,” The Journal of Machine Learning Research, vol. 18, no. 1, pp. 7947– 7980, 2017.

[9] A. U. Rahman and G. Ghatak, “A beam-switching scheme for resilient mm-wave communications with dynamic link blockages,” in IEEE WiOpt, 2019.

[10] Q. Zhu and V. Y. Tan, “Thompson sampling algorithms for mean-variance bandits,” arXiv preprint arXiv:2002.00232, 2020.

[11] E. Contal, D. Buffoni, A. Robicquet, and N. Vayatis, “Parallel gaussian process optimization with upper confidence bound and pure exploration,” in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2013, pp. 225–240.

[12] W. R. Thompson, “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples,” Biometrika, vol. 25, no. 3/4, pp. 285–294, 1933.

[13] O.-C. Granmo, “Solving two-armed bernoulli bandit problems using a bayesian learning automaton,” International Journal of Intelligent Computing and Cybernetics, 2010.

[14] S. Agrawal and N. Goyal, “Analysis of thompson sampling for the multi⍰armed bandit problem,” in Conference on learning theory, 2012, pp. 39–1.

[15] A. Garivier and E. Moulines, “On upper-confidence bound policies for non-stationary bandit problems,” arXiv preprint arXiv:0805.3415, 2008.

[16] O. Besbes et al., “Stochastic multi-armed-bandit problem with non⍰stationary rewards,” in Advances in neural information processing sys⍰tems, 2014, pp. 199–207.

[17] C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag, “Multi⍰armed bandit, dynamic environments and meta-bandits,” 2006.

[18] J. Y. Yu and S. Mannor, “Piecewise-stationary bandit problems with side observations,” in Proceedings of the 26th

annual international conference on machine learning, 2009, pp. 1177–1184.

[19] Y. Cao, Z. Wen, B. Kveton, and Y. Xie, “Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit,” in The 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 418–427.

[20] O. Besbes et al., “Optimal exploration-exploitation in a multi-armed bandit problem with non-stationary rewards,” Stochastic Systems, vol. 9, no. 4, pp. 319–337, 2019.

[21] J. Mellor and J. Shapiro, “Thompson sampling in switching environ�ments with bayesian online change detection,” in Artificial Intelligence and Statistics, 2013, pp. 442–450.

[22] V. Srivastava, P. Reverdy, and N. E. Leonard, “Surveillance in an abruptly changing world via multiarmed bandits,” in 53rd IEEE Con�ference on Decision and Control. IEEE, 2014, pp. 692–697.

[23] G. Ghatak, “A change-detection based thompson sampling framework for non-stationary bandits,” IEEE Transactions on Computers, pp. 1–1, 2020.

[24] P. Auer, P. Gajane, and R. Ortner, “Adaptively tracking the best bandit arm with an unknown number of distribution changes,” in Conference on Learning Theory, 2019, pp. 138–158.

[25] S. S. Villar et al., “Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges,” Statistical science: a review journal of the Institute of Mathematical Statistics, vol. 30, no. 2, p. 199, 2015.

[26] K. Jahanbakhsh, “Applying multi-armed bandit algorithms to computa�tional advertising,” arXiv preprint arXiv:2011.10919, 2020.

[27] V. W. Berger and Y. Zhou, “Kolmogorov–smirnov test: Overview,” Wiley statsref: Statistics reference online, 2014.

[28] P. Massart, “The tight constant in the dvoretzky-kiefer-wolfowitz in�equality,” The annals of Probability, pp. 1269–1283, 1990.

[29] S. Shalev-Shwartz and S. Ben-David, Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.

[30] N.-N. Dao, T.-T. Nguyen, M.-Q. Luong, T. Nguyen-Thanh, W. Na, and S. Cho, “Self-calibrated edge computation for unmodeled time-sensitive iot offloading traffic,” IEEE Access, vol. 8, pp. 110 316–110 323, 2020.

[31] A. U. Rahman, G. Ghatak, and A. De Domenico, “An online algorithm for computation offloading in non-stationary environments,” IEEE Com�munications Letters, vol. 24, no. 10, pp. 2167–2171, 2020.

[32] F. Black and R. Litterman, “Global portfolio optimization,” Financial analysts journal, vol. 48, no. 5, pp. 28–43, 1992.

[33] D. Gawlik. (2017) New York Stock Exchange: S&P 500 companies historical prices with fundamental data. [Online]. Available: https: //kaggle.com/dgawlik/nyse

[34] H. A. Al-Zeaud, “Modelling and forecasting volatility using arima model,” European Journal of Economics, Finance & Administrative Science, vol. 35, pp. 109–125, 2011.

[35] S. J. Taylor and B. Letham, “Forecasting at scale,” The American Statistician, vol. 72, no. 1, pp. 37–45, 2018.