NON-STATIONARY BANDIT CHANGE DETECTION-BASED THOMPSON SAMPLING ALGORITHM

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 - *Many rigorous mathematical approaches and optimum solutions to the stationary multi-armed bandit (MAB) paradigm may be found in the literature. However, the MAB issue is famously difficult to analyse for a non-stationary environment, i.e. when the reward distribution fluctuates over time. There are two main methods that have been suggested to combat non-stationary bandit problems: i) passively adaptable methods, which can be easily analysed, and ii) actively adaptive methods, which monitor their surroundings and make adjustments in response to any changes they discover. Researchers have responded by developing new bandit algorithms that build on previously established methods, such as the sliding-window upper-confidence bound (SW-UCB), the dynamic upper-confidence bound (d-UCB), the discounted upper-confidence bound (D-UCB), the discounted Thompson sampling (DTS), etc. For this reason, we focus on the piecewise stationary setting, in which the reward distribution is held constant for some period of time before changing at some arbitrary moment. For this context, we offer the TS-CD family of change-detection-based, actively-adaptive TS algorithms. In specifically, a Poisson arrival process is used to mimic the non-stationary environment, which adjusts the reward distribution with each new arrival. We use the Kolmogorov-Smirnov test (KS-test) and the Anderson-Darling test (AD-test) as Goodness-of-fit tests to identify the shift. When TS detects a shift, it either updates the algorithm's parameters or penalises previous successes. We have conducted experiments on edge-control of i) multi-connectivity1 and ii) RAT selection in a wireless network to evaluate the efficacy of the proposed method. We have compared the TS-CD algorithms to other bandit algorithms including D-UCB, discounted Thompson sampling (DTS), and change detection-based UCB (CD-UCB), all of which are optimised for dynamic situations. We demonstrate the higher performance of the proposed TS-CD in the studied applications via comprehensive simulations.*

Key Words: Decision-making, Contextual bandits, Probability distribution, Empirical performance, Convergence

1. INTRODUCTION

The multi-armed bandit (MAB) framework is a classical problem in decision-making that involves making a series of choices among several alternatives (or "arms") with uncertain rewards. The problem gets its name from the idea of a gambler standing in front of a row of slot machines, each with a different payout rate. The gambler must choose which machines to play and in what order while trying to maximize their overall payout.

In the MAB framework, each arm corresponds to a choice or action that can be taken, and the rewards associated with each arm are random variables with unknown distributions. The goal is to learn which arm(s) have the highest expected reward, based on a limited number of trials or samples. The challenge is to balance exploration (trying out different arms to learn their reward distributions) with exploitation (selecting the arm with the highest expected reward based on the current knowledge).

Several algorithms have been developed for solving the MAB problem, including the epsilon-greedy algorithm, the upper confidence bound (UCB) algorithm, and Thompson sampling. These algorithms differ in their strategies for balancing exploration and exploitation.

The MAB framework has applications in many fields, including online advertising, clinical trials, and recommendation systems. In each case, the problem involves making decisions with uncertain outcomes and limited resources. By using MAB algorithms to optimize decision-making, it is possible to improve performance and increase efficiency in a wide range of settings.

1.1. THOMPSON SAMPLING ALGORITHM

Thompson Sampling is a popular algorithm for solving the multi-armed bandit problem. It is a Bayesian approach that balances exploration and exploitation by selecting actions according to their probabilities of being optimal, based on the observed data. The basic idea of Thompson Sampling is to model the reward distribution for each arm as a probability distribution, such as a beta distribution for binary rewards or a Gaussian distribution for continuous rewards. The algorithm then maintains a posterior distribution over the parameters of each distribution, which is updated as new data is observed.

At each iteration, Thompson Sampling samples a set of parameters from the posterior distribution for each arm and selects the arm with the highest expected reward based on the sampled parameters. This balancing of exploration and exploitation is achieved by the probabilistic nature of the algorithm, where it randomly selects the arms and estimates their rewards based on the current belief about the distribution of rewards.

The key advantage of Thompson Sampling is its ability to naturally handle uncertainty in the reward distributions, without requiring explicit assumptions about the form of the distributions. This makes it well-suited to problems where the reward distributions are complex or unknown.

Thompson Sampling has been shown to have strong theoretical guarantees in the context of regret analysis, which measures the performance of an algorithm relative to the best possible reward that could have been achieved. It has also been shown to outperform other popular algorithms like epsilon-greedy and UCB in a variety of settings, including online advertising and recommendation systems.

2. ALGORITHMS FOR NON-STATIONARY ENVIRONMENT TOP OF FORM

In a non-stationary environment, the reward distributions of the arms can change over time, which presents a challenge for standard multi-armed bandit algorithms. Here are some algorithms that are designed to handle non-stationary environments:

Exp3: The Exp3 algorithm is an extension of the classic multi-armed bandit algorithm that uses exponential weighting to balance exploration and exploitation. It also includes a learning rate parameter that can be tuned to adapt to changes in the reward distribution over time.

UCB-V: The UCB-V (Upper Confidence Bound with Variance) algorithm is a modification of the UCB algorithm that incorporates the variance of the reward distribution as an additional factor in the decision-making process. This makes it more robust to changes in the reward distribution over time.

CUSUM-UCB: The CUSUM-UCB algorithm combines the UCB algorithm with the CUSUM (cumulative sum) control chart method from statistical process control. This allows it to detect changes in the reward distribution and adjust the exploration rate accordingly.

Dynamic Thompson Sampling: Dynamic Thompson Sampling is a variant of the Thompson Sampling algorithm that updates the prior distribution over the reward distribution parameters over time. This allows it to adapt to changes in the reward distribution and maintain a balance between exploration and exploitation.

Sliding-Window UCB: Sliding-Window UCB is a modification of the UCB algorithm that uses a sliding window of the most recent rewards to estimate the mean and variance of the reward distribution. This allows it to adapt to changes in the reward distribution over time and avoid overfitting to outdated data.

These algorithms are just a few examples of the many approaches that have been developed for solving the multi-armed bandit problem in non-stationary environments. Each algorithm has its strengths and weaknesses, and the choice of algorithm will depend on the specific problem and constraints.

3. CHANGE DETECTION BASED ON UCB

Change Detection Based UCB (CD-UCB) is a variant of the Upper Confidence Bound (UCB) algorithm that is designed to handle non-stationary environments where the reward distributions of the arms can change over time.

The CD-UCB algorithm uses a change detection mechanism to detect when a change in the reward distribution has occurred. When a change is detected, the algorithm switches to an exploration phase to collect more data and update its estimate of the new reward distribution.

The change detection mechanism used by CD-UCB is based on the Page-Hinkley test, which is a statistical test for detecting changes in a data stream. The test looks for a significant increase or decrease in the mean of the reward distribution, relative to a baseline mean.

When a change is detected, the algorithm switches to an exploration phase, where it samples each arm with a probability proportional to its uncertainty. This allows the algorithm to collect more data on the new reward distribution and update its estimates of the mean and variance.

Once enough data has been collected, the algorithm switches back to the exploitation phase and selects the arm with the highest upper confidence bound, as in the standard UCB algorithm. The CD-UCB algorithm continues to monitor the reward distribution and switch between exploration and exploitation phases as necessary.

CD-UCB has been shown to outperform other change detection algorithms and standard UCB algorithms in several experiments, including online advertising and recommendation systems. However, the performance of the algorithm depends on the choice of parameters and the properties of the data stream.

4. EXPERIMENTAL SETUP

Using a piecewise stationary approach, we will talk about the problem statement that was presented to us and establish the foundation for our future work environment. In the next paragraph, we will discuss the details of how this will take place. To put it another way, we will have our working environment ready. In addition to this, we will talk about the change detection methods that we use in our work and the operational ideas behind them.

Estimation of the Mean Amount of Change

Mean Estimated Change Detection, often known as MECD, is an example of a change detection method that is typically implemented inside the framework of the multi-armed bandit problem. It is intended to identify changes in the mean of the reward distribution of an arm and to adjust the exploration rate appropriately in response to these detections.

The MECD method keeps an estimate of both the mean and the variance of the reward distribution for each arm of the experiment. These estimates are kept in sync with one another. Next, it evaluates the accuracy of the current estimate of the mean by contrasting it to a reference value that is initially based on the actual value that represents the mean of the reward distribution.

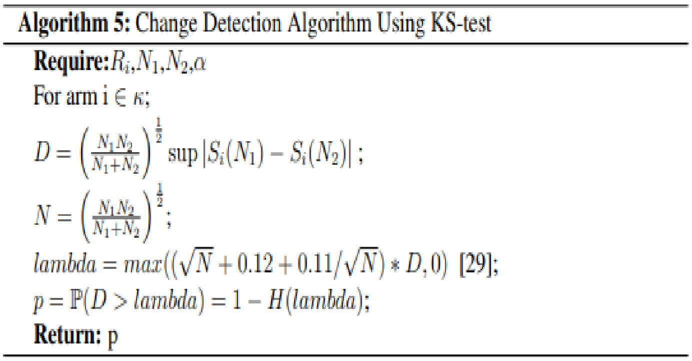
The algorithm concludes that there has been a change after a change has happened when the difference between the most recent estimate of the mean and the reference value is greater than a specific threshold. This causes the algorithm to raise the exploration rate. This allows the algorithm to gather further data on the revised reward distribution and to bring its estimate of the mean and variance up to date.

After collecting a sufficient amount of data, the algorithm will transition back to the exploitation phase, at which point it will choose the limb that has the greatest estimated mean. The MECD algorithm will continue to keep an eye on the reward distribution and will make any required adjustments to the exploration pace.

MECD is effective at detecting changes in the mean of the reward distribution, particularly in circumstances where the changes are gradual rather than rapid. This is especially true in situations where the changes are more likely to occur gradually. However, the algorithm might not be as good at detecting changes in the other properties of the reward distribution, such as the variance or the shape of the distribution. This could be a limitation of the algorithm.

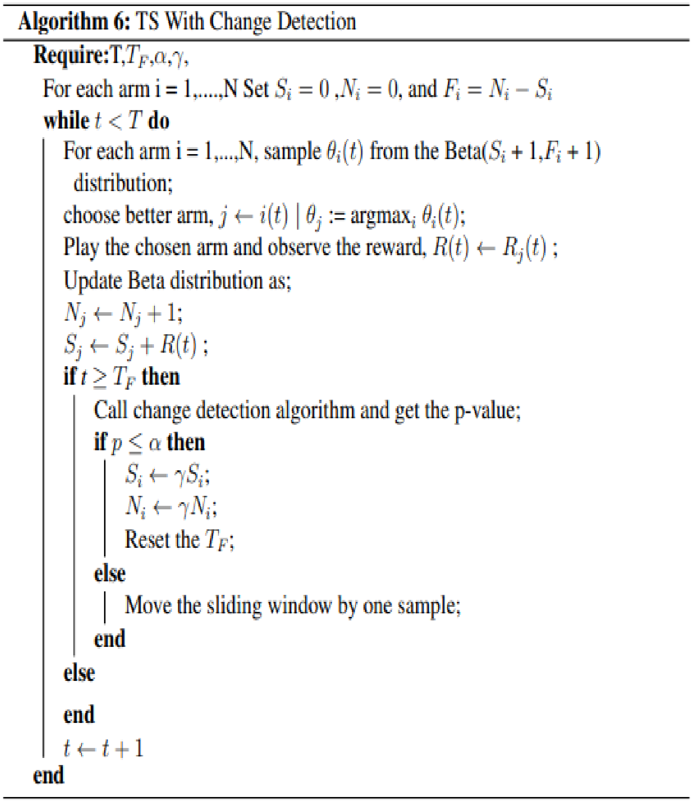
5. ALGORITHM FOR DETECTING CHANGES

When the current time, t, is greater than TF, the CD technique encourages the very immediate commencement of its execution process. This ensures that the technique may continue to function normally. This guarantees that an adequate quantity of data is available, from which an accurate approximation of the empirical distributions may be derived. After that, to locate the shift, we first build the empirical distributions by using the historical records of payments made for each of the arms. This allows us to determine the shift. Because of this, we can pinpoint the exact position of the shift.



Algorithm 5: Change Detection Algorithm Using KS-test

```
Require:$R_i$,$N_1$,$N_2$,$\alpha$
For arm i $\in \kappa$;
$D = \left(\frac{N_1 N_2}{N_1+N_2}\right)^{\frac{1}{2}} \sup |S_i(N_1) - S_i(N_2)|$ ;
$N = \left(\frac{N_1 N_2}{N_1+N_2}\right)^{\frac{1}{2}}$;
$lambda = max((\sqrt{N} + 0.12 + 0.11/\sqrt{N}) * D, 0)$ [29];
$p = \mathbb{P}(D > lambda) = 1 - H(lambda)$;
Return: p
```



Algorithm 6: TS With Change Detection

```
Require:T,$T_F$,$\alpha$,$\gamma$,
For each arm i = 1,....,N Set $S_i = 0$ ,$N_i = 0$, and $F_i = N_i - S_i$
while $t < T$ do
    For each arm i = 1,...,N, sample $\theta_i(t)$ from the Beta($S_i$ + 1,$F_i$ + 1)
     distribution;
    choose better arm, $j \leftarrow i(t) \mid \theta_j :=$ argmax$_i$ $\theta_i(t)$;
    Play the chosen arm and observe the reward, $R(t) \leftarrow R_j(t)$ ;
    Update Beta distribution as;
    $N_j \leftarrow N_j + 1$;
    $S_j \leftarrow S_j + R(t)$ ;
    if $t \geq T_F$ then
        Call change detection algorithm and get the p-value;
        if $p \leq \alpha$ then
            $S_i \leftarrow \gamma S_i$;
            $N_i \leftarrow \gamma N_i$;
            Reset the $T_F$;
        else
            Move the sliding window by one sample;
        end
    else

    end
    $t \leftarrow t + 1$
end
```

6. SIMULATION AND RESULT ANALYSIS

It is very required for the connection of user equipment (UE) in 5G and other wireless applications to be synchronized to the time-varying dynamics of both the environment and the user equipment. This is the case regardless of whether the applications are wired or wireless. It makes no difference whether the apps are wired or wireless; this is true in either scenario. This is still the case even if the individual software does not enable wireless communication. When seen from this perspective, it does not make a difference whether the application in issue is wired or wireless. Both approaches have their benefits and drawbacks. In any scenario, the end consequence is the same. To go further, it is necessary to satisfy these requirements in some way. It is impossible to do anything else. We are going to test the method that was provided for the scenario that involves a large number of connections, and we are also going to replicate the number of connections to the typical UE for the MAB framework to be able to determine the appropriate number of arms for the solution. These tests and replications are going to be carried out to ensure that the solution is effective. For us to be able to identify the number of connections that are involved in the scenario that includes a significant number of connections, these tests and replications are going to be carried out.

6.1. RAT SELECTION IN WIRELESS NETWORKS

When selecting a RAT, it is essential to consider the network infrastructure and the devices' capabilities. For example, older devices may only support 2G or 3 G networks, while newer devices may support 4G or 5G networks. Similarly, the network infrastructure may not support all RATs, so it is crucial to choose a RAT that is compatible with the network infrastructure.

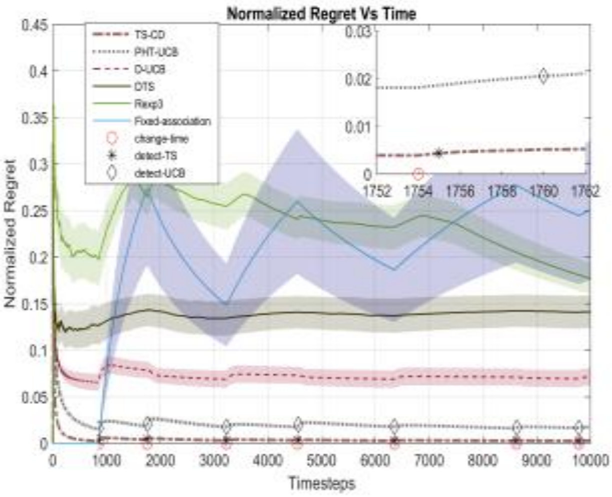


Figure -1: Time-averaged regret for various algorithms.

This illustrates that the TS-CD technique that is now being contemplated is capable of monitoring all of the changes that take place in an environment that is dynamic with a reasonable degree of ease provided that the required hyper-parameter tweaking is carried out.

6.2. MULTI-CONNECTIVITY

To get things moving in the right direction, we are going to start by defining a performance metric that is unique to us. Things will get off to a good start as a result of this. As a consequence of this, we will have the ability to launch the procedure. We are going to refer to the average effective throughput (AET) as the following as a performance metric with n connections since it is as follows. This is a suitable amount of connections taking into account the fact that it has that many connections already. Throughout the process of calculating the AET, the following factors were taken into account, which explains why this is the case:

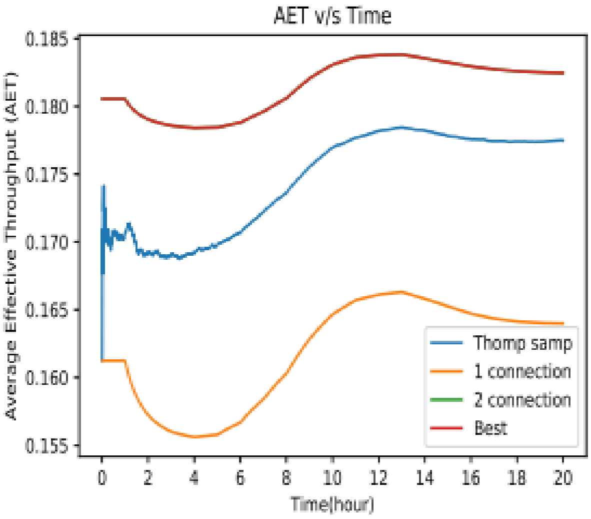


Figure-2: AET The effectiveness of the suggested algorithm in comparison to conventional, fixed-association techniques.

In the beginning, the performance of the scenario with two connections is higher than that of the case with one connection, but by the time t = 9h has elapsed, the performance of the case with one connection has overtaken that of the scenario with two connections. This is something that is brought to our notice every once in a while. This is something that we can verify for ourselves and something that we can see for ourselves as well. The algorithm that was devised, known as TS-CD, was able to accurately monitor the alteration, and it would then choose the limb that has the potential to provide the highest number of advantages.

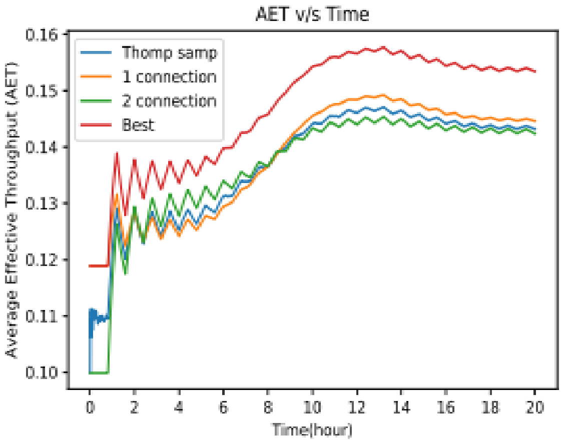


Figure-3: AET The effectiveness of the suggested algorithm in comparison to a variety of static association strategies.

7. CONCLUSION

Thesis conclusion? Check these. Summarises the thesis's key findings and questions. Show how the research advanced the field. Discuss research flaws. Examine the study's context. Consider the research's practicality.

The non-stationary two-armed bandit problem was addressed using change-detection-based Thompson Sampling. This method is TS-CD. It was the non-stationary two-armed bandit. The non-stationary two-armed bandit research evaluated this strategy. This ensured environmental compatibility. TS-CD's immobility regime time was estimated. Our computation employed the lowest time limitation feasible. Duration of immobility limited this. We assessed the stable regime's duration, which restricts us. Finally, TS-CD may achieve asymptotic optimality by limiting alterations. Asymptotic optimality. Limitations force the best answer. The edge-of-a-wireless-network RAT selection challenge tested our technique. We can test the method. TS-CD outperformed max power band selection and dynamic bandit. TS-CD wins both. Benchmarks comparing TS CD performance. TS-CD surpassed its competition. We graded each strategy's effectiveness to achieve this. Regular CDs and TS-performance Discs were compared. The non-stationary two-armed bandit problem was addressed using change-detection-based Thompson Sampling. This method is TS-CD. It was the non-stationary two-armed bandit. The non-stationary two-armed bandit research evaluated this strategy. This ensured environmental compatibility.

We evaluated TS-CD's stationary period before detecting a change. That timeframe's worst-case scenario was estimated. Long-term immobility required this limitation. Stable regime forecasts limit this. Finally, we demonstrate that the TS-CD approach, which reduces iterations to examine modifications, may reach asymptotic optimality. Asymptotic perfection is possible. The algorithm refines towards the best result. We tested network-edge RAT selection. This aids in technique evaluation. TS-CD outperforms industry-standard max power band selection and previously published bandit algorithms under different situations. TS-CD beats CD. TS CD findings are compared to numerous other metrics. TS-CD's higher performance enabled this. Technique effectiveness was rated. against this, TS-performance Discs were compared to CDs.

REFERENCE

[1] Villar, Sofía S., Jack Bowden, and James Wason. "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 30.2 (2015): 199.

[2] Buccapatnam, Swapna, et al. "Reward maximization under uncertainty: Leveraging side-observations on networks." The Journal of Machine Learning Research 18.1 (2017): 7947-7980.

[3] Rahman, Aniq Ur, and Gourab Ghatak. "A Beam-Switching Scheme for Resilient mm-Wave Communications With Dynamic Link Blockages." Workshop on Machine Learning for Communications, WiOpt, IEEE. 2019.

[4] Contal, Emile, et al. "Parallel Gaussian process optimization with upper confidence bound and pure exploration." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2013.

[5] Thompson, William R. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples." Biometrika 25.3/4 (1933): 285-294.

[6] Chapelle, Olivier, and Lihong Li. "An empirical evaluation of thompson sampling." Advances in neural information processing systems. 2011.

[7] Raj, Vishnu, and Sheetal Kalyani. "Taming non-stationary bandits: A Bayesian approach." arXiv preprint arXiv:1707.09727 (2017).

[8] Garivier, Aurélien, and Eric Moulines. "On upper-confidence bound policies for switching bandit problems." International Conference on Algorithmic Learning Theory. Springer, Berlin, Heidelberg, 2011.

[9] Liu, Fang, Joohyun Lee, and Ness Shroff. "A change-detection based framework for piecewise-stationary multi-armed bandit problem." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[10] Gupta, Neha, Ole-Christoffer Granmo, and Ashok Agrawala. "Thompson sampling for dynamic multi-armed bandits." 2011 10th International Conference on Machine Learning and Applications and Workshops. Vol. 1. IEEE, 2011.

[11] Hartland, Cédric, et al. "Multi-armed bandit, dynamic environments and metabandits." (2006).

[12] Raj, Vishnu, and Sheetal Kalyani. "Taming non-stationary bandits: A Bayesian approach." arXiv preprint arXiv:1707.09727 (2017).

[13] Ghatak, Gourab, Antonio De Domenico, and Marceau Coupechoux. "Coverage analysis and load balancing in HetNets with millimeter wave multi-RAT small cells." IEEE Transactions on Wireless Communications 17.5 (2018): 3154- 3169.

[14] Bai, Tianyang, and Robert W. Heath. "Coverage and rate analysis for millimeter-wave cellular networks." IEEE Transactions on Wireless Communications 14.2 (2014): 1100-1114.

[15] Bai, Tianyang, Rahul Vaze, and Robert W. Heath. "Analysis of blockage effects on urban cellular networks." IEEE Transactions on Wireless Communications 13.9 (2014): 5070-5083.

[16] Zhang, Xinchen, and Martin Haenggi. "A stochastic geometry analysis of intercell interference coordination and intra-cell diversity." IEEE Transactions on Wireless Communications 13.12 (2014): 6655-6669.