Non Stationary Bandits with Periodic Variation

Extended Abstract

Titas Chakraborty*
Indian Institute of Technology Bombay
Mumbai, India
titas0602@gmail.com

Parth Shettiwar*
Indian Institute of Technology Bombay
Mumbai, India
parthshettiwar@g.ucla.edu

ABSTRACT

In numerous real-world scenarios, we encounter periodic patterns in the dynamics of non-stationary data. Unfortunately, current approaches to addressing non-stationary bandit problems overlook the valuable potential offered by the presence of periodicity. In response, we introduce SW-PUCB, a novel sliding window algorithm explicitly designed to exploit periodicity in bandit arms, surpassing the performance of the conventional UCB approach when dealing with perfectly periodic bandit environments. Recognizing that perfect periodicity is seldom encountered in real-world setting, we further present SW-NPUCB, another sliding window algorithm tailored to data exhibiting near-periodic characteristics. Lastly, we demonstrate the practical efficacy of our algorithms through comprehensive experimentation, on synthetically generated data. By bench-marking against existing non-stationary bandit techniques, we emphasize the superiority of our approaches.

KEYWORDS

Reinforcement Learning, Multi-armed bandits, Non-stationary bandits, Periodic environment

ACM Reference Format:
Titas Chakraborty and Parth Shettiwar. 2024. Non Stationary Bandits with Periodic Variation: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

In real world applications of non-stationary multi armed bandits, there are many cases where the non-stationarity in the true means of the arms is periodic in nature. For instance, in network selection, data usage and network quality follow periodic patterns daily [6]. Metrics such as google searches and number of users in a store also follow predictable patterns on a yearly basis, if not exactly periodic patterns [7]. We can exploit this periodicity in non-stationary multi-armed bandits to get better estimates of the true arms means

There have been numerous papers tacking the non-stationary multi-armed bandit problem. [1] introduces two approaches towards solving this problem, a sliding window approach and a discounted approach. More recently, there have been several other approaches [3]. In this paper, we attempt to extend existing approaches for solving non-stationary multi armed bandits for a case where the distributions of the rewards for all arms is periodic with period P. Our contributions from this work are:

- We introduce a new setting in non-stationary bandits where the true means of arms can vary in a periodic fashion.
- We propose algorithms SW-PUCB and SW-NPUCB for the perfectly periodic and nearly periodic settings respectively and show their regret bounds.
- We evaluate our algorithms performance in a continuously varying simulated environment where they outperform all existing previous works adapted in our settings

2 PROBLEM FORMULATION

We consider a non-stationary multi armed bandit problem where there are k arms. To simplify our proofs, we consider that the reward of each arm is bounded in the range $[0, 1]$ with mean $r_{i,t}$ for arm i at time step t. The means for all the arms are non-stationary. T is the horizon. We analyse 2 cases:

- **Perfectly Periodic**: We assume that $r_{i,t}$ is periodic with a period of P time steps. More formally, $r_{i,t+P} = r_{i,t}$ is satisfied $\forall i$ and $\forall t$. We define a piece-wise budget B such that the means for all arms i satisfy the condition $|r_{i,t+1} - r_{i,t}| \leq B$ for all timesteps t.
- **Nearly Periodic**: We allow the means of arms to vary between periods. Therefore, we have two budgets B_1 and B_2. $\sum_{i=1}^{i=T-P} |r_{i,t+P} - r_{i,t}| \leq B_2$ is satisfied $\forall i$. Here B_2 is called the periodic budget for our problem. Also a piece-wise budget B_1, like the previous setting, holds true for $\forall i$ and $\forall t$.

3 PERFECTLY PERIODIC BANDITS

In this section we analyse the case when means of arms are Perfectly periodic with a known period P. We propose an algorithm, namely **Sliding Window Periodic UCB (SW-PUCB)**. The sliding window algorithm was first studied by [1]. Here we do an intuitive modification to it in our periodic setting by considering window length rewards at each time step, P steps apart. In addition to this, we also consider a naive UCB estimate, which considers rewards only at time steps apart by multiple of periods. The arm picked at each time step is the maximum of the minimum of the above 2 UCB estimates. Intuitively, we improve over the naive UCB algorithm by using an additional UCB estimate obtained from more timesteps. SWUCB is described in 1

*Equal contribution

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Algorithm 1 SW-PUCB

```
Input: $P$, $B$, number of arms $k$, horizon $T$, window $w$
Initialization:- $R(a,t) = 0$, $N(a,t) = 0$, $UC(a,t) = 1$, $LC(a,t) = -1$
For first $k$ periods, pull arm $i$ for the $i^{th}$ period
for t = k*P to T do
    $t_r = t\%P$
    $n_{at} = \sum_{l=t_r-w}^{l=t_r+w} N(a, l\%P)$
    $d(i,t) = \min(i\%P - t\%P, t\%P - i\%P)$
    $e(a,i) = \min(d(i,t)B\ ,\ UC(a,t_r) - LC(a,i))$
    $\hat{x}_{at,1} = \frac{\sum_{i=t_r-w}^{i=t_r+w}(R(a,i)+e(a,i)N(a,i))}{\sum_{i=t_r-w}^{i=t_r+w}N(a,i)}$
    $\hat{x}_{at,2} = \frac{R(a,t_r)}{N(a,t_r)}$
    $v_1(a) = \hat{x}_{at,1} + \sqrt{\frac{3ln(t)}{2n_{at}}}$
    $v_2(a) = \hat{x}_{at,2} + \sqrt{\frac{3ln(t)}{2N(a,t_r)}}$
    $a_{opt} = \underset{a}{\operatorname{argmax}}\ \min(v_1(a), v_2(a))$
    Pull arm $a_{opt}$ and obtain reward $r$
    $N(a_{opt}, t_r) = N(a_{opt}, t_r) + 1$
    $R(a_{opt}, t_r) = R(a_{opt}, t_r) + r$
    $UC(a_{opt}, t_r) = \frac{R(a_{opt},t_r)}{N(a_{opt},t_r)} + \sqrt{\frac{3ln(t)}{2N(a_{opt},t_r)}}$
    $LC(a_{opt}, t_r) = \frac{R(a_{opt},t_r)}{N(a_{opt},t_r)}$ - $\sqrt{\frac{3ln(t)}{2N(a_{opt},t_r)}}$
end for
```

The regret bound for the SW-PUCB algorithm can be derived with relative ease as follows:

$$R(T) \leq \sum_{i=0}^{i=P-1} \sum_{a \neq a^*} \frac{6ln(T)}{\Delta_{a,i}} + \frac{k\pi^2}{2}$$

The above expression implies that our algorithm cannot perform worse than a naive UCB approach where every P^{th} sample is considered by more than a constant factor.

4 NEARLY PERIODIC BANDITS

In this section, we explore nearly periodic bandits, where across periods the means of arms can change, bounded by a periodic budget B_2. We propose a sliding window algorithm, namely **Sliding Window Nearly Periodic UCB (SW-NPUCB)**.

To account for the variation across periods, we propose to use another window w_2, which would determine the number of periods from which rewards would be taken for UCB estimate. The w_1 window would still be present to account for the piece-wise budget. The SW-NPUCB Algorithm is described in 2. Refraining from delving into a detailed proof, we ascertain that the total regret of SW-NPUCB is bounded as :-

$$R(T) \leq 2w_1w_2B_1 + 2w_1B_1 + 2w_2B_2 + 8T\sqrt{6ln(T)}\frac{\sqrt{|A|}}{\sqrt{w_1w_2}} + \frac{k\pi^2}{3}$$

Optimizing for w_1 and w_2, we can get the order of regret $R(T)$ as:

$$R(T) = O(T^{\frac{2}{3}}|A|^{\frac{1}{3}}B_1^{\frac{1}{3}} + T^{\frac{1}{3}}|A|^{\frac{1}{6}}B_1^{\frac{1}{6}}B_2^{\frac{1}{2}})$$

Algorithm 2 SW-NPUCB

```
Input: $P$, number of arms $k$, horizon $T$, window sizes $w_1$ and $w_2$
Initialization:- $R(a, t_r, [\frac{t}{P}]) = 0$, $N(a, t_r, [\frac{t}{P}]) = 0$
for t = 1 to T do
    $t_r = t\%P$
    $\hat{x}_{at} = \frac{\sum_{k=\max([\frac{t}{P}]-w_2+1,0)}^{k=[\frac{t}{P}]} \sum_{l=t_r-w_1+1}^{l=t_r+w_1-1} R(a,l\%P,k)}{\max(1,\sum_{k=\max([\frac{t}{P}]-w_2+1,0)}^{k=[\frac{t}{P}]} \sum_{l=t_r-w_1+1}^{l=t_r+w_1-1} N(a,l\%P,k))}$
    $n_{at} = \sum_{k=\max([\frac{t}{P}]-w_2+1,0)}^{k=[\frac{t}{P}]} \sum_{l=t_r-w_1+1}^{l=t_r+w_1-1} N(a, l\%P, k)$
    $c_{at} = \min(1, \sqrt{\frac{3ln(t)}{2n_{at}}})$
    $v(a) = \hat{x}_{at} + c_{at}$
    $a_{opt} = \underset{a}{\operatorname{argmax}}\ v(a)$
    Pull arm $a_{opt}$ and obtain reward $r$
    $N(a_{opt}, t_r, [\frac{t}{P}]) = N(a_{opt}, t_r, [\frac{t}{P}]) + 1$
    $R(a_{opt}, t_r, [\frac{t}{P}]) = R(a_{opt}, t_r, [\frac{t}{P}]) + r$
end for
```

5 RESULTS

In this section, we evaluate our algorithms in a continuously varying environment where we vary the means of arms in a sinusoidal fashion with a period $P = 100$ and plot the cumulative regret over time. For the nearly periodic case, we add some random uniform noise over these waves so that they are not perfectly periodic. For baselines comparison, for perfectly periodic case, the naive UCB algorithm is adapted by taking rewards at period P time steps apart. For nearly periodic case, we adapt REXP3 [2], DTS [4], SWUCB [8] and D-LinUCB [5] algorithms with their optimal gamma and window values. We observe that, our algorithms clearly outperform the baselines implying that it is imperative to leverage periodicity in a dataset to get optimal performance (minimize regret)

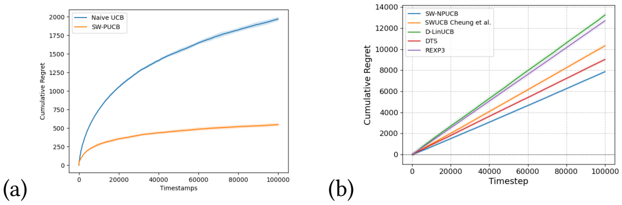

Figure 1: Comparison of (a) SW-PUCB with naive UCB, and (b) SW-NPUCB with other non-stationary bandit algorithms

6 CONCLUSION AND FUTURE WORK

In this work, we have introduced a new framework in non stationary bandits by considering periodic variation in the rewards. and proposed 2 novel algorithms, SW-PUCB and SW-NPUCB

There is great scope for expansion in our work. We used a simple rectangular window in our work. Others may wish to experiment with different window shapes. In some cases period might be unknown to the agent or might vary after some duration. Handling such cases would also improve the practicability.

REFERENCES

[1] Eric Moulines Aurélien Garivier. [n.d.]. On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems.

[2] Omar Besbes, Yonatan Gur, and Assaf Zeevi. 2019. Optimal Exploration–Exploitation in a Multi-armed Bandit Problem with Non-stationary Rewards. *Stochastic Systems* 9 (10 2019). https://doi.org/10.1287/stsy.2019.0033

[3] Yuan Jiang Zhi-Hua Zhou Peng Zhao, Lijun Zhang. [n.d.]. A Simple Approach for Non-stationary Linear Bandits.

[4] Vishnu Raj and S. Kalyani. 2017. Taming Non-stationary Bandits: A Bayesian Approach. *ArXiv* abs/1707.09727 (2017).

[5] Yoan Russac, Claire Vernade, and Olivier Cappé. 2019. Weighted Linear Bandits for Non-Stationary Environments.

[6] Seth Gilbert Shunhao Oh, Anuja Meetoo Appavoo. [n.d.]. Periodic Bandits and Wireless Network Selection.

[7] Weiyu Yan Stefano Tracà, Cynthia Rudin. [n.d.]. Regulating Greed Over Time in Multi-Armed Bandits.

[8] Ruihao Zhu Wang Chi Cheung, David Simchi-Levi. [n.d.]. Learning to Optimize under Non-Stationarity.