Change Detection Based Thompson Sampling Algorithm For Non-Stationary Bandits

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

M.Tech

BY KUNWAR ZAID

Electronics and Communication Engineering

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI
NEW DELHI– 110020

Abstract

The stationary multi-armed bandit (MAB) framework is a well-studied problem in literature, with many rigorous mathematical treatments and optimal solutions. However, for a non-stationary environment, i.e., when the reward distribution changes over time, the MAB problem is notoriously difficult to analyze. In general, to address non-stationary bandit problems, researchers have proposed two approaches: i) passively adaptive techniques, that are analytically tractable, or ii) actively adaptive techniques that keep track of the environment and adapt as soon as changes are detected. Consequently, researchers have come up with variants of bandit algorithms that are based on classical solutions, e.g., sliding-window upper-confidence bound (SW-UCB), dynamic UCB (d-UCB), discounted UCB (D-UCB), discounted Thompson sampling (DTS), etc. In this regard, we consider the piecewise stationary environment, where the reward distribution remains stationary for a random time and changes at an unknown instant. We propose a class of change-detection based, actively-adaptive, TS algorithms for this framework named TS-CD. In particular, the non-stationary in the environment is modeled as a Poisson arrival process, which changes the reward distribution on each arrival. For detecting the change we employ i) mean-estimation based methods, and ii) Goodness-of-fit tests, namely the Kolmogorov-Smirnov test (KS-test) and the Anderson-Darling test (AD-test). Once a change is detected, the TS algorithm either refreshes the parameters, or discounts the past rewards. To assess the performance of the proposed algorithm, we have tested it for edge-control of i) multi-connectivity[1] and ii) RAT selection in a wireless network. We have compared the TS-CD algorithms with other bandit algorithms that are designed for non-stationary environments, such as D-UCB, discounted Thompson sampling (DTS) and change detection based UCB (CD-UCB). With extensive simulations, we establish the superior performance of the proposed TS-CD in the considered applications.

[1]This work is under minor revision in Elsevier Physical Communication.

Contents

List of Figures **iii**

1 Introduction **1**
1.1 Stochastic Bandits . . . 2
1.2 Algorithms For The Stationary Bandits . . . 2
1.2.1 ϵ - greedy . . . 2
1.2.2 Upper Confidence Bound (UCB) . . . 3
1.2.3 Thompson Sampling . . . 3
1.3 Algorithms for Non-Stationary Environment . . . 4
1.3.1 Change detection Based UCB . . . 4
1.3.2 Discounted Upper Confidence bound . . . 5
1.3.3 Sliding Window Upper Confidence Bound . . . 6
1.3.4 Discounted Thompson Sampling . . . 6

2 Experimental Setup **8**
2.1 Problem Statement: The Two Armed-Bandit Setting . . . 8
2.2 Piecewise-Stationary Environment . . . 8
2.3 Change Detection Methods . . . 8
2.3.1 Mean Estimated Change Detection . . . 8
2.3.2 Goodness Of Fit Test Based Change Detection . . . 10

3 Proposed Algorithm **14**
3.1 Change Detection Algorithm . . . 14
3.2 Thompson Sampling With Change Detection . . . 15

4 Case Study **17**
4.1 System Model . . . 17
4.2 Characterization Of Mean Rewards . . . 18

5 Simulation Results **20**
5.1 RAT Selection In Wireless Networks . . . 20
5.2 Multi-Connectivity . . . 21

Bibliography **23**

List of Figures

5.1 Time averaged regret for different algorithms. 20
5.2 AET Performance of the proposed algorithm as compared to static association schemes with Fixed β. 21
5.3 AET Performance of the proposed algorithm as compared to static association schemes with varying β. 22

List of Illustrations

Chapter 1

Introduction

The multi-armed bandit (MAB) framework is a class of problems in online learning and sequential decision making. Since its introduction, it has found many applications in the fields of reinforcement learning [28], online recommendation systems [27], clinical trials [1], computational advertisement systems [2], and wireless communications [3]. The classical MAB problem models the exploration-exploitation trade-off inherent in sequential decision problems. In MAB framework at each step a learning agent pulls an arm of a K-armed bandit based on its past observations, and receives a reward accordingly. Each arm is characterized by a unknown reward distribution and the rewards are independent and identically distributed (iid).
The objective of a learning agent or algorithm is to maximize the total expected reward, or to minimize the expected regret over time horizon T, which is defined as the expectation of the difference of total reward obtained and that of the highest expected reward. To facilitate this, the algorithm keeps track of each arm, so that according to the past history of rewards, either it can select the current best arm or explore other arms, which is an exploration/exploitation dilemma. Stationary multi-armed bandit in which the reward distribution is time invariant is very well studied in the literature. Algorithms like upper confidence bound (UCB) [4], have been proven to perform optimally. Thompson sampling, which was first discussed in [5] has also got some interesting results. For some applications such as online advertising Thompson sampling's performance is far better than other algorithms [6].
However, when the reward distribution is non-stationary, some researches have proposed the idea of discounting the past rewards to make the system adaptive to the dynamic changes [7], [8]. Garivier and Moulines [8] have presented a scenario where environment is stationary for fixed duration of time and changes abruptly at unknown time. They have analysed the theoretical upper bounds of regret for the discounted UCB and sliding window UCB. Liu $et\ al.$ [9] has also considered the piece-wise stationary environment and presented the change detection based framework for multi-armed bandit problems and studied the class of change detection based UCB policies (CD-UCB). They have used cumulative sum and Page-Hinkely statistical test (PHT) as the change detection algorithm.

Gupta $et\ al.$ [10], has proposed Dynamic Thompson sampling. Hartland $et\ al.$ [11] has also considered dynamic bandits and abrupt changes in the environment and proposed an algorithm called Adapt-EvE. It also uses change point detection technique to detect any abrupt change in the environment and to detect the change it also uses Page-Hinkely test. It utilizes a meta bandit formulation for exploration-exploitation dilemma once the change is detected. For the non-stationary environment, Raj and Kalyani [12] have proposed an algorithm based on Bayesian bandits. Their algorithm- Discounted Thompson sampling (DTS), discounts the effect of past rewards.

1.1 Stochastic Bandits

Reinforcement learning is a field that uses the training information to evaluate the actions taken rather than instructing by giving correct actions, this is one of the important feature that distinguish reinforcement learning from other types of learning. This is the reason of exploration, the requirement of explicit search for good behaviour. Purely evaluative feedback is the indication of the goodness of an action taken, it may be the best action or the worst possible. The evaluative feedback problem that we will explore is the $K-$armed bandit problem/ Multi-armed bandits (MAB). Bandit problem was first introduced by William R. Thompson in 1993 [25]. In early days, medical trials were run blindly, without adapting the treatment allocations on the fly as the drug appears more or less effective. The name comes from the study of Frederick Mosteller and Robert Bush in 1950s [25]. They decided to study animal learning and ran trials on mice and then on humans. They took a T-shaped maze and put a piece of food at one end each time, unknown to the mice. Each time the mice has to decide to which side it should go. A similar experiment was performed with the humans, they were faced with a dilemma to pull either left or right arm of a 'two-armed machine'. Each time when an arm is pulled, the machine returns a random payoff with the distribution of payoffs unknown to the human player, the machine was called 'two-armed bandit'. The MAB problem is one of the topics which is studied for decades in statistics, operations research, electrical engineering, computer science, and economics.

A stochastic bandit consist of a set of arms $\kappa = \{1, 2, ...K\}$ with distributions $\langle D_1, ...D_k \rangle$ and mean $\langle \mu_1, ...\mu_k \rangle$ associated with each arm $I_t \in \kappa$. The player interacts with the environment for T rounds, T is the horizon. In each round $t \in \{1, 2, ...T\}$, the player chooses an arm/action $I_t \in \kappa$ and receives a reward $X_t(I_t)$ from distribution D_{I_t}. The player's goal is to maximize the total expected reward or the total expected regret, that depends on the actions of the player. Regret is the amount of deficit suffered by the player by not selecting the optimal action in each round, defined as

$$R(T) = \mathbb{E}\left[\sum_{t=1}^{T} \left(X_t(i_t^*) - X_t(i_t)\right)\right].$$

where, i_t^* is the optimal arm.

1.2 Algorithms For The Stationary Bandits

To gain more insight of the bandit problem, we will go through with some of the algorithms that are present in the literature. There are many algorithms for the stationary bandits, but we will only explore ϵ - greedy, upper confidence bound (UCB), and Thompson Sampling (TS).

1.2.1 ϵ - greedy

The ϵ- greedy algorithm is very simple. At each time step $t = 1, 2, ...$, the algorithm selects the arm which has the highest estimated reward up-to time t with

probability $1-\epsilon$ and explores other arms with probability ϵ. Mathematically

$$p_i(t+1) = \begin{cases} 1-\epsilon+\epsilon/k; & i = \arg\max_{i=1,2,..K} \mu_j^*(t) \\ \epsilon/k & otherwise \end{cases}$$

the performance of $\epsilon-$ greedy algorithm depends on the choice of the exploration factor ϵ.

1.2.2 Upper Confidence Bound (UCB)

The principle of optimism in the face of uncertainty is the basis of the upper confidence bound algorithm, which means that one should act as if the environment is as acceptable as credibly possible. In context of bandits, optimism means to assign a value called upper confidence bound to each arm using the information obtained so far, which is an overestimate of the unknown mean with the high probability. The others arms will be explored only if the upper confidence bound of these arms is greater than that of the optimal arm, which in turn is larger than the mean of the optimal arm [26].
The idea of UCB algorithm is that it tracks the number of times each arm is played up-to time t, denoted by $N_i(t)$. It is ensured that each arm is played once initially, and after that at each time-step t selects an arm j as follows:

$$j(t) = \arg\max_{i=1,2,...K} \left(\hat{\mu}_i + c\sqrt{\frac{\ln t}{N_i}} \right)$$

$\hat{\mu}_i$ is the estimate of mean reward, and $c > 0$ is a constant that controls the degree of exploration. Therefore, the quantity which is maximized is a sort of upper-bound on the true value, and the quantity c determining the confidence level. each time an arm i is selected: N_i increases, and in return decreases the uncertainty, as it appears in the denominator. If some other arm is selected, the uncertainty increases, as t in the denominator increases and N_i remains the same.

1.2.3 Thompson Sampling

Thomson sampling - also known as posterior sampling and probability matching, was proposed by Thomson in 1993, for the two-armed bandit problem for clinical trial. Thomson sampling is considered among the Bayesian bandits, approach which involve a prior over a problem instance. For the sake of understanding, we will consider the Bernoulli bandit problem, i.e., reward, r_t is either 0 or 1, and for arm i probability of success is μ_i, μ_i is the unknown expected reward of arm i [18]. The rewards are obtained immediately after the arm is played and are i.i.d. Also the observed rewards are independent of the plays of the other arms. Let the prior distribution be the Beta distribution. It is a convenient choice due to its conjugacy property. The beta distribution forms a family of continuous probability distribution on the interval $(0, 1)$. The pdf of the beta prior with parameters α,β is defined as:

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where, $\alpha > 0$, and $\beta > 0$. The mean of the $Beta(\alpha, \beta)$ is $\frac{\alpha}{\alpha+\beta}$; it is evident from the distribution that, higher the value of α,β, higher will be the concentration of the curve around the mean. The posterior update is very simple under the beta prior, if on pulling an arm there is success, then the update will be $Beta(\alpha + r_t, \beta)$, if there is a failure, then it will be $Beta(\alpha, \beta + r_t)$. The algorithm starts with the prior distribution, $Beta(1, 1)$, which is a uniform distribution on $(0, 1)$. Let $S_i(t)$ be the number of success of arm i up-to time t and $F_i(t)$ be the number of failures. Let $N_i(t) = S_i(t) + N_i(t)$ be the total number of pulls of arm i. The algorithm updates the distributions on μ_i as $Beta(S_i(t) + 1, N_i(t) - S_i(t) + 1$, and samples $\theta_i(t)$ from this posterior distribution and selects an arm $i(t)$ according to the rule:

$$i(t) = \arg\max_i \theta_i(t)$$

and observe reward r_t.

1.3 Algorithms for Non-Stationary Environment

The multi-armed bandit problem is widely studied in the literature for the stationary environment. However, in real world problems, this assumption of stationary environment does not hold, therefore it is necessary to study the problem of non-stationary bandits. To understand the non-stationary bandits, we will discuss some of the work that is present in the literature. We will discuss some algorithms such as, sliding-window upper-confidence bound (SW-UCB), change-detection based UCB (CD-UCB), discounted UCB (D-UCB), and discounted Thompson sampling (DTS). Some other algorithms are also present, but we will only briefly go through with these.

1.3.1 Change detection Based UCB

Change detection based framework under piece-wise stationary environment is presented by Fang Liu. *et al*, in [9]. Authors have studied change detection based UCB algorithm, that detects the change in the environment and restarts the bandit algorithm. The CD-UCB consists of change detection algorithm and a bandit algorithm. For change detection authors have used cumulative sum method and Page Hinkley test (Hinkley 1971), and UCB is used as the bandit algorithm. The CD-UCB algorithm as presented in [9] is shown in Algorithm1.

CD($\cdot$,$\cdot$) is the change detection algorithm which takes arm index i and observation $X_t(i)$ as input, and on detecting a change, returns 1. For more details on change detection algorithm that authors have used, one can refer to [9]. The change detection algorithm combined with the UCB is shown in algorithm 1, known as CD-UCB. Now, we will describe some notations and some equations of UCB algorithm that authors have used. τ_i is the last time CD(i,$\cdot$) alarms a change and restarts for arm i before time t. $N_i(t)$ is the number of valid observations for arm i up to time t, and n_t is the total number of observations. α is the tuning parameter, $\bar{X}_t(i)$ is the sample

Algorithm 1: CD-UCB

Require: T,α, and an algorithm CD($\cdot$,$\cdot$)
Initialize $\tau_i = 1$, $\forall_i$
while $t < T$ **do**
  Update according to equations (1.1 - 1.3) ;
  Play arm I_t and observe $X_t(I_t)$;
  if $CD\left(I_t, X_t(I_t)\right) == 1$ **then**
    $\tau_{I_t} = t + 1$;
    reset CD(I_t,$\cdot$)
  else
  end
end

average and, $C_t(i)$ is the confidence padding term. Therefore,

$$\bar{X}_t(i) = \sum_{s=\tau_i}^{t} \frac{X_s(i)}{N_t(i)} \mathbb{1}_{\{I_s=i\}} \qquad N_t(i) = \sum_{s=\tau_i}^{t} \mathbb{1}_{\{I_s=i\}}$$

$$C_t(i) = \sqrt{\frac{\xi \log n_t}{N_t(i)}} \qquad n_t = \sum_{i=1}^{K} N_t(i)$$

$$I_t = \begin{cases} \arg\max_{i \in \kappa} \left(\bar{X}_t(t) + C_t(i)\right), & w.p \quad 1-\alpha \\ i, & \forall i \in \kappa \quad w.p \quad \frac{\alpha}{k} \end{cases}$$

To gain more insight of CD-UCB policy, one can refer to [9].

1.3.2 Discounted Upper Confidence bound

In this section we will discuss another algorithm for the non-stationary environment, proposed by Kocsis and Szepesv´ari (2006). They have proposed a discounted variant of the UCB algorithm, with discounting factor $\gamma \in (0, 1)$. For the instantaneous expected reward, policy constructs an UCB $\bar{X}_t(\gamma, i) + C_t(\gamma, i)$, and the discounted empirical average is

$$\bar{X}_t(\gamma, i) = \frac{1}{N_t(\gamma, i)} \sum_{s=1}^{t} \gamma^{t-s} X_s(i) \mathbb{1}_{\{I_s=i\}}$$

$$N_t(\gamma, i) = \sum_{s=1}^{t} \gamma^{t-s} \mathbb{1}_{\{I_s=i\}}$$

and the discounted padding function is

$$C_t(\gamma, i) = \sqrt{\frac{\xi \log n_t(\gamma)}{N_t(\gamma, i)}}, \qquad n_t(\gamma) = \sum_{i=1}^{K} N_t(\gamma, i)$$

From the above equations one can see that if $\gamma = 1$, the discounted version of UCB is same as UCB algorithm. using above equation, discounted UCB (D-UCB) is defined in Algorithm 2. To estimate the instantaneous expected reward, the D-UCB algorithm averages past rewards in a way that, it assigns more weight to the recent rewards.

Algorithm 2: Discounted UCB

for t from 1 to K, play arm
    $I_t = t$;
for t from K+1 T, play arm
    $I_t = \arg\max_{1 \leq i \leq K} \bar{X}_t(\gamma, i) + C_t(\gamma, i)$

1.3.3 Sliding Window Upper Confidence Bound

The sliding window UCB (SW-UCB) is proposed by Garivier and Moulines (2008). In the D-UCB algorithm, we saw that all the past rewards were considered for taking the average, with a discounting factor that assigns more weight to the recent reward. In SW-UCB, averages are computed for a fixed time period. At time t, for averaging, instead of using all the past rewards with a discount factor, SW-UCB relies on a local empirical average of the observed, using only the last τ plays. For the instantaneous expected reward, the algorithm constructs an UCB $\bar{X}_t(\tau, i) + C_t(\tau, i)$, and the local empirical average reward is given by:

$$\bar{X}_t(\tau, i) = \frac{1}{N_t(\tau, i)} \sum_{s=t-\tau+1}^{t} X_s(i) \mathbb{1}_{\{I_s = i\}}$$

and the padding function is

$$C_t(\tau, i) = B\sqrt{\frac{\xi \log(n_t \wedge \tau)}{N_t(\tau, i)}}$$

where $t \wedge \tau$ is the minimum of t and τ, and ξ is a constant. The SW-UCB is defined in Algorithm 3.

Algorithm 3: Sliding Window UCB

for t from 1 to K, play arm
    $I_t = t$;
for t from K+1 T, play arm
    $I_t = \arg\max_{1 \leq i \leq K} \bar{X}_t(\tau, i) + C_t(\tau, i)$

1.3.4 Discounted Thompson Sampling

It is a variant of Thompson sampling algorithm proposed in [12], for restless bandits. As discussed in the earlier in section of stationary bandits, Thompson sampling

maintains a prior distribution, and samples from this prior distribution for selecting the arm to play. Beta distribution is used as prior distribution, having parameters α and β, as already discussed in the section of Thompson sampling. The key idea of discounted Thompson sampling (dTS) is that it uses a discounting factor, $\gamma \in (0, 1)$, to discount the past rewards, and also, it increases the variance of the unexplored arms systematically. Another feature of dTS is that, while modifying the variance of the arms, the mean is kept constant between the plays. The mean is modified only for the arm which is played. The dTS is defined in Algorithm 4. dTS differs with

Algorithm 4: Discounted Thompson sampling

Require:T,T_F,α,γ,
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;
  $F_j \leftarrow \gamma F_j + (1 - R_t)$;
  $S_j \leftarrow \gamma S_j + R(t)$;
  $F_{i \neq j} \leftarrow \gamma F_{i \neq j}$;
  $S_{i \neq j} \leftarrow \gamma S_{i \neq j}$;
end

TS in a way that, it discounts the past values of S_i and F_i before updating with the current rewards. Note that, when $\gamma = 1$, dTS is same as the TS. Authors have also derived the probability of choosing a sub-optimal arm for dTS, and how the variance is changed because of the discounting factor. To gain more insight of dTS, one can refer to [12].

Chapter 2

Experimental Setup

In this section we will describe the problem statement and will also set-up piecewise stationary environment for our work. Further, we will also explain the change detection methods that we have used in our work.

2.1 Problem Statement: The Two Armed-Bandit Setting

Let $\kappa = \{1, 2\}$ be the set of arms. Let $t = \{1, 2, ...T\}$ be the time instants when the decision maker makes the decision and T is the time horizon. At each time step the player chooses an arm $I_t \in \kappa$ and obtains a reward $X_t(I_t)$. We assume that the each arm $I_t \in \kappa$ has Gaussian distribution[1], i.e., $X_t(I_t) \sim \mathcal{N}(\mu_t(I_t), \sigma^2)$. The arm $I_t \in \kappa$ has unknown, non-stationary mean ($\mu_t(I_t)$) and known, fixed variance (σ^2). We have assumed that the mean μ_j, $j \in \kappa$ remains constant for some time and changes at unknown time-instants T_{C_i}, where i = 1,2,..., with T_{C_0} assumed to be at t = 0.

2.2 Piecewise-Stationary Environment

We consider a model where the reward process of the arms is non-stationary on the whole, but stationary on intervals. The reward distribution changes arbitrarily and at arbitrary time steps, otherwise remains stationary.
We have assumed that the change time instants, T_{C_i}, follows the Poisson arrival process with parameter, λ_C, and the inter-arrival time is exponentially distributed:

$$\mathbb{P}(T_{C_{i+1}} - T_{C_i} \leq k) = 1 - exp(-\lambda_C k)$$

We choose the value of λ_C in such a way that the reward distribution is stationary for required time and we get enough number of samples for the goodness of fit test.

2.3 Change Detection Methods

2.3.1 Mean Estimated Change Detection

We will consider the two armed-bandit setting as described in the previous section. For the mean estimated method, we will use the results derived by G. Ghatak in [23] [2]. Author has made an assumption that the reward distribution remains stationary for $T_F = T_N + n_T$ time steps after every time a change occurs where n_T is the number of time-slots after a change, required to detect a change, and T_N is the minimum number of time-slots for which the MAB framework remains stationary after the detection of the change. The change will be detected when the mean of test sequence (μ_{test})

[1]This assumption is made in regards of the use case, i.e., RAT selection strategy. In general it is not necessary that the reward distribution is Gaussian. Our change detection algorithm will work even for other distributions.

[2]This work is under review

differs from the mean of the estimate sequence $(\hat{\mu_i})$ by more than Δ_C. Let's say that the change occurs at time T_{C_i} and detected at time $n = T_{D_i} > T_{C_i}$. For evaluating μ_{test}, out of the n_T samples of μ_{test}, let n_1 samples $(X_1, X_2, ...X_{n_1})$ comes from the distribution $X_i \sim \mathcal{N}(\mu_i(T_{C_i} - 1), \sigma^2)$ and n_2 samples $(Y_1, Y_2, ..., Y_N)$ from $Y_i \sim \mathcal{N}(\mu_i(T_{C_i} + 1), \sigma^2)$. The mean of arm I_t changes from $\mu_i(T_{C_i} - 1)$ to $\mu_i(T_{C_i} + 1)$ with condition that $|\mu_i(T_{C_i} - 1) - \mu_i(T_{C_i} + 1)| \geq \Delta_m$, the minimum difference between the mean rewards of the same arm across the time slots. Also, the author has assumed that, without loss of generality $\mu_i(T_{C_i} + 1) < \mu_i(T_{C_i} - 1)$ .Therefore the detection criteria is:

$$\left| \frac{1}{n_T} \left(\sum_{p=n-n_T}^{n-n_T+n_1} X_p + \sum_{n-n_T+n_1}^{n_1} Y_p \right) - \frac{1}{N_i} \left(\sum_{q=n-n_T-N_i}^{n-n_T} X_q \right) \right| \geq \Delta_C$$

Now, the points where mean estimated based change detection algorithm will fail are:

- E_1: The estimate of the mean of the original distribution is not accurate $(|\hat{\mu_1} - \mu_1| > \epsilon)$. Let T_n be the minimum number of samples for an accurate estimation, as mentioned earlier.
- E_2: The change detection framework is not able to detect the change due to less number of samples from the test distribution. Let n_T be the number of samples from the test distribution for efficient estimation.
- E_3: Two consecutive changes in the arm occur too often for the detection to keep track. This results in the upper bound on change parameter λ_C.

The mathematical characterization of the above events is done in [23] and the results are as follows:

- Under E_1, the minimum number of plays of TS-CD framework for estimate of mean, μ_i of best arm to be well-localized with probability greater than $1 - p_{loc}$ is

$$T_N = \frac{-40}{\Delta_\mu^2} \mathcal{W} \left(-\exp \left(\frac{-40}{\Delta_\mu^2} \left(\frac{1}{\epsilon} \ln \left(\frac{1}{p_{loc}} \right) - \frac{48}{\Delta_\mu^4} \right) \right) \frac{\Delta_\mu^2}{40} \right)$$

- Under E_2, given μ_i is well-localized, for a false alarm probability $\mathcal{P}_F$, to limit the probability of failure of change detection to $\mathcal{P}_M$, the number of samples required in test set are

$$n_T = \frac{1}{\Delta_m} \left(\sqrt{\ln \frac{1}{\mathcal{P}_M}} + \sigma Q^{-1}(\mathcal{P}_F) \right)$$

- Under E_3, to limit the probability of frequency of change to p_{change}, the bound on the value of λ_C is

$$\lambda_C \leq \frac{1}{n_T + T_N} \ln \left(\frac{1}{1 - p_{change}} \right)$$

And,

$$\Delta_C = \frac{\sigma Q^{-1}(\mathcal{P}_F)}{\sqrt{n_T}} - \epsilon$$

The proof follows the procedure in [23].

2.3.2 Goodness Of Fit Test Based Change Detection

The gof test measures how well the observed values of data fits a distribution. It is used to test whether the two samples are coming from the same distribution or from different distributions. In this section, we will present Kolmogorov-Smirnov test (KS-test) and Anderson-Darling test (AD-test). We will compare both of them and will present an analysis on the number of samples required to detect the change.

Kolmogorov Smirnov Test

The Kolmogorov-Smirnov test was introduced by Kolmogorov(1933,1941), and Smirnov (1939). It is a non-parametric hypothesis test that measures that an i.i.d. sample $X_1, X_2, ...X_n$ comes from a particular distribution or not. We will first discuss the one-sample KS-test and will extend this to two-sample test. Consider an i.i.d. sample $X_1, X_2, ...X_n$ that comes from some unknown distribution $\mathbb{P}$. The objective is to test the hypothesis that $\mathbb{P}$ is equal to a particular distribution $\mathbb{P}_0$, that is, to test the hypothesis:

$$H_0 : \mathbb{P} = \mathbb{P}_0$$
$$H_1 : \mathbb{P} \neq \mathbb{P}_0$$

There are many advantages of KS-test, such as:

- It is sensitive to the shape of a distribution because it can detect differences everywhere along the axis.
- It is applicable for the small sample sizes as well.
- It can detect very small changes in the distribution.
- It does not care if the sample size of the two data is unequal.

Now, let $F(x) = \mathbb{P}(X \leq x)$ denotes the c.d.f of random variable x. Now, let's define the empirical c.d.f as

$$F_n(x) = \mathbb{P}_n(X \leq x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

According to the law of large numbers, for any fixed point $x \in \mathbb{R}$

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \rightarrow \mathbb{E}I(X \leq x) = \mathbb{P}(X \leq x) = F(x)$$

which means that

$$\sup_x |F_n(x) - F(x)| \to 0$$

i.e. maximum difference between empirical c.d.f. and the true c.d.f. approaches zero in probability. There are two important theorems that we will discuss now

Theorem 1. *If F(x) is continuous then the distribution of*

$$\sup_x |F_n(x) - F(x)|$$

does not depend on F.

Proof. The proof follows the procedure in [25] □

Another important result we will use is

$$\mathbb{P}\left(\sqrt{n}\sup_x |F_n(x) - F(x)| \leq t\right) \to H(t) = 1 - 2\sum_1^\infty (-1)^{i-1} e^{-2i^2 t}$$

where $H(t)$ is the c.d.f. of Kolmogorov-Smirnov distribution, and

$$D_n = \sup_x |F_n(x) - F(x)|$$

is the KS-statistics.
Now, we will present the KS-statistics for the two sample KS-test, that we have used in our change detection algorithm. Let us consider that the sample $X_1, X_2, ...X_n$ of size n and has a distribution with c.d.f. $F(x)$ and another sample $Y_1, Y_2, ...Y_m$ of size m has distribution with c.d.f. $G(x)$, and we want to test whether the two samples are coming from same distribution or different. Let $F_n(x)$ and $G_m(x)$ be the empirical c.d.f.s, then the KS-statistics is:

$$D_{nm} = \left(\frac{nm}{m+n}\right)^{\frac{1}{2}} \sup_x |F_n(x) - G_m(x)|$$

The KS-statistics for two sample test also follows the results presented above. Now let's derive the KS-statistics for the Gaussian distribution. let us consider that

$$f_1(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\left(\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)\right)$$
$$f_2(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\left(\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)\right)$$

and the corresponding c.d.f.s. are

$$F_1(x) = \int_{-\infty}^{x} f_1(x)dx$$
$$F_2(x) = \int_{-\infty}^{x} f_2(x)dx$$

respectively. We know from the fundamental theorem of calculus that the derivative of a function:

$$F(x) = \int_0^t f(x)dx$$

is

$$\frac{d}{dx}\int_0^x f(t)dt = f(x)$$

Therefore ${F_1}'(x) = f_1(x)$ and ${F_2}'(x) = f_2(x)$. Now, the KS-statistics is the maximum over x of the difference of the two c.d.f.s, therefore to maximize we will take the derivative and will equate it to zero to find the value of x that maximize the statistics, as shown below:

$$\Rightarrow \frac{d}{dx}\left|F_1(x) - F_2(x)\right| = 0$$

$$\Rightarrow f_1(x) = f_2(x)$$

$$\Rightarrow \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\left(\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)\right) = \frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\left(\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)\right)$$

on solving we will get

$$x = \frac{\mu_1 + \mu_2}{2}$$

Since, we do not know the exact mean and variance, i.e. μ_1,μ_2 and σ_1, σ_2 are unknown, we will use their estimates to find the statistics. Therefore

$$x = \frac{\hat{\mu_1} + \hat{\mu_2}}{2}$$

Now the KS-statistics is

$$\left|F_1(x) - F_2(x)\right|_{x=\frac{\hat{\mu_1}+\hat{\mu_2}}{2}}$$

and we know that

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-t^2}{2}\right)$$

and

$$G(x) = 1 - Q(x)$$

where, $G(x)$ is the c.d.f. of standard Normal distribution. From the above two results we can say that

$$F_1(x) = 1 - Q\left(\frac{x-\hat{\mu_1}}{\hat{\sigma_1}}\right)$$

and

$$F_2(x) = 1 - Q\left(\frac{x - \hat{\mu_2}}{\hat{\sigma_2}}\right)$$

$$\Rightarrow |F_1(x) - F_2(x)|_{x=\frac{\hat{\mu_1}+\hat{\mu_2}}{2}} = \left|Q\left(\frac{\hat{\mu_1} - \hat{\mu_2}}{2\sigma}\right) - Q\left(\frac{\hat{\mu_2} - \hat{\mu_1}}{2\sigma}\right)\right|$$

Let's assume that $\sigma_1 = \sigma_2 = \sigma$, and using the property of $Q - function$, $Q(x) = 1 - Q(-x)$

$$D = \left|2Q\left(\frac{\hat{\mu_1} - \hat{\mu_2}}{2\sigma}\right) - 1\right|$$

Anderson Darling Test

The KS-test Works effectively when the distributions differ in center. But, when the change in the two distributions is around the tails, KS-test does not perform well. Therefore, it is required to have a test which can detect smaller changes at any point along the distribution. The Anderson-Darling (AD) test was developed in 1952 by T.W. Anderson and D.A. Darling (Anderson and Darling, 1952). We will first understand the one sample AD-test. Let $\left\{X_{(1)} < X_{(2)} < ..... < X_{(n)}\right\}$, be the observed data of sample size n, and let $F(X)$ be the underlying cumulative distribution to which the sample data is considered. The one sample AD statistics is

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(\ln(x_{(i)}) + \ln(1 - (x_{(n+1-i)})))$$

The null hypothesis is that, the observed data comes from the underlying distribution $F(X)$. The null hypothesis is rejected or the change is detected when the critical value AD_α is less than the AD-statistics for a given value of α.
The two-sample test was introduced by Darling (1957) and Pettitt (1976). The two-sample Ad statistics is

$$AD = \frac{1}{mn}\sum_{i=1}^{m+n}\left(N_i Z_{(m+n-ni)}\right)^2 \frac{1}{iZ_{(m+n-i)}}$$

where $Z_{(m+n)}$ is the combined and ordered samples X_m and Y_n of size m and n respectively. N_i is the number of observations in X_m that are less than or equal to the i_{th} observation in $Z_{(m+n)}$. The change is detected i.e., X_m and Y_n comes from the different distributions when the AD statistics is greater than the critical value. Further , Anderson Darling test is also generalized for k-samples, but we will not discuss here. The AD test have all the advantages of KS-test, discussed in the previous section. Apart from that, AD-test has two more advantages. The AD test is highly sensitive towards the tails of the distribution, in which KS-test has less sensitivity, and AD-test can detect very small changes even for large sample size [19].

Chapter 3

Proposed Algorithm

In this chapter we will present the change detection algorithm using goodness of fit test and will use that algorithm for our main algorithm, "change detection based Thompson sampling using goodness of fit test".

3.1 Change Detection Algorithm

First we will propose the change detection (CD) algorithm using KS-test in Algorithm 1. For the change detection we will require, time duration T_F for which the environment is stationary, history of rewards R_i of arm i $\in \kappa$. Let N_1 be the number of samples before T_F and N_2 be the test samples, i.e., number of samples against which the change detection is performed. Note that, some samples from time before T_F are also considered in N_2 to make a good estimate.
Let $S_i(N_1)$ and $S_i(N_2)$ be the empirical c.d.f.s of arm i before and after time T_F respectively, and let $F_i(N_1)$ and $F_i(N_2)$ be the corresponding true distributions.
The CD algorithm starts just after time t is greater than T_F, so that we have enough number of samples to make the good estimate of empirical distributions. First we form the empirical distributions with the history of rewards for each arm to detect the change.

Assumption 1. *For all the arms, change is happening at the same unknown time instants.*

Note that, for the ease of simulations, we have assumed that all the arms are changing at the same time, this may not true in general. Due to this assumption we only have to apply CD algorithm to any one arm, but if all the arms are changing at different time it will be applied to all arms in parallel.
Using these empirical distributions, we will calculate the KS-statistics, which follows the KS-distribution. And finally we will calculate the p-value to compare with acceptable significance level α. If $p - value < \alpha$, then we will say that the change is detected, i.e., both the reward samples are coming from different distributions.

Algorithm 5: Change Detection Algorithm Using KS-test

Require: R_i,N_1,N_2,α
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

3.2 Thompson Sampling With Change Detection

We propose TS-CD in Algorithm 2 for the non-stationary environment. In TS-CD, first we initialize the parameters of the Beta distribution S_i and N_i for i $\in \kappa$ with 0. At each time-step t we sample from the Beta distribution for all the arms and play the arm which returns the highest sample and obtain the reward $R_i(t)$ for that arm. Consequently we update the posterior distribution of the arm played.
We have already defined T_F in the above sections, to detect the change, at each time step we will see whether t is greater than T_F or not, if condition is true than we call the change detection algorithm, i.e., Algorithm 1. CD algorithm will return a p-value, which will be compared with the given acceptable significance level α. If the change is detected then parameters of the posterior distribution will be reset according to the equation,

$$S_i \leftarrow \gamma S_i$$
$$N_i \leftarrow \gamma N_i$$

Note that, if the Value of γ is 0, then the parameters are completely refreshed. γ is the hyper-parameter, which we have to tune to get the best results.
If the change is not detected than we will move the sliding window by one sample, which means that we will consider the newest sample in the empirical distribution and will discard the oldest sample, keeping number of samples that correspond to the distributions fixed.

Algorithm 6: TS With Change Detection

Require:T,T_F,α,γ,
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

Chapter 4

Case Study

Since beginning, mobile technologies are designed to serve the purpose of the end users. As the world is evolving, there is a tremendous increase in the demand for high data-rates. Societies are becoming more and more data-centric and automated and this requires transfer of much greater amounts of data, at much higher speeds. Transmission in high frequency-range, such as mm-Wave is one promising solution. However, mm-Wave transmission suffers a lot due to path-loss and blockages [17]. It is important that the existing cellular architectures must complement the first generation of mm-Wave access point (AP) deployment. However, the user equipment association (UE) to the different APs and to different available frequency bands need to be dynamic due to the environment dynamics, such as vehicular and human blockages.

Rahman $et\ al$ [3] have studied a windowed TS based mm-Wave beam-forming scheme which switches between two APs based on the blockage conditions. However, they have not studied how to choose the duration of the window. In this chapter, we will analyse a band-switching scheme modelled as two armed-bandit problem, and study the efficacy of the TS-CD algorithm.

Another application that we will explore is of multi-connectivity or macro-diversity. In multi-connectivity, a single user equipment (UE) connects to multiple base stations simultaneously and increase the cellular coverage of the UEs [20], [21]. Macro-diversity has many advantages, although its effects on the network throughput have not studied properly in the literature. Connecting to multiple BSs of a single user might degrade the network throughput due to over-provisioning of the resources. In 5G and other wireless applications, the connection of UEs is required to be synced to the time-varying dynamics of the environment and user equipment. We will test the proposed algorithm for the case of multi-connectivity and will model the number of connections to the typical UE as the arms of the MAB framework.

4.1 System Model

We consider a wireless system consisting of APs on the the two dimensional Euclidean plane. APs are located as points of a homogeneous Poisson point process (PPP)ϕ, with intensity λ. Without loss of generality, we consider that the user is located at the origin and it connects to AP with the highest downlink received power. To provide ad-hoc coverage and enhanced-data rates, APs are assumed to operate in two RATs, sub-6GHz band and the mm-Wave Band [13]. For simplification, let us denote RAT with r, where $r \in \{m, s\}$ stands for mm-Wave and sub-6GHz respectively. Let $v \in \{L, N\}$ be the visibility for line-of-sight (LOS) and non-line-of-sight (NLOS), respectively. We assume that received power at the typical user from a AP at a distance d fromthe user is given by $K_r P_r x^{-\alpha_{rv}}$, where α_{rv} is the path loss exponent for RAT r and visibility state v, K_r is the path loss exponent for RAT r and P_r is the transmit power from AP in RAT r. $\sigma^2_{N,r}$ is the noise power in RAT r. The

received power in mm-Wave takes the advantage of directional antenna gain of the transmitter and receiver. The user and the serving base station are assumed to be aligned and the interfering base stations are assumed to be randomly oriented with respect to the typical user. Here, we assume a model where the product of transmitter and receiver antenna gain, G, takes on the values a_k with probabilities b_k as given in table 1 of [14]. Let the maximum value of G be G_0.
We assume the channel visibility (v) in the sub-6GHz and the mm-Wave to be the same for a given AP, because, the probability of blockage of a signal is independent of the frequency band, it mainly depends on blockage process [15]. We further categorize ϕ, into either LOS NLOS processes: ϕ_L and ϕ_N, respectively, due to the blockage conditions. The intensity of these modified processes are given by $p(x)\lambda$ and $1-p(x)\lambda$, respectively, where $p(x)$ is the probability of a AP at a distance x to be in LOS with respect to the typical user. For tractability we assume the following function [13]:

$$p(x) = \begin{cases} 1; & x \leq d \\ 0 & x > d \end{cases}$$

That is, AP within the distance d from the user will be considered as LOS and beyond d will be NLOS. where, d is the LOS ball radius [13]. To study the performance of signal to interference and noise ratio (SINR), first, the path loss process is reformulated as one dimensional process, $\phi'_{vr} = \left\{\xi_{vr,i} : \xi_{vr,i} = \frac{\|x\|^{\alpha_{vr}}}{K_r P_r}, x_i \in \phi_v\right\}$, $v \in \{L, N\}, r \in \{s, m\}$. The processes ϕ'_{vr} are non-homogeneous with intensities calculated as below.

Lemma 2. *The intensity measures of LOS and NLOS path loss processes, ϕ'_{Lr} and ϕ'_{Nr} are:*

$$\Lambda'_{LR}(0,x) = \begin{cases} \pi\lambda \left(K_r P_r\right)^{\frac{2}{\alpha_{Lr}}} x^{\frac{2}{\alpha_{Lr}}}, & x < \frac{d^{\alpha_{Lr}}}{K_r P_r} \\ \pi\lambda d^2, & x > \frac{d^{\alpha_{Lr}}}{K_{vr} P_r} \end{cases}$$

$$\Lambda'_{NR}(0,x) = \begin{cases} 0, & x < \frac{d^{\alpha_{Nr}}}{K_{vr} P_r} \\ \pi\lambda \left(\left(K_r P_r x\right)^{\frac{2}{\alpha_{Nr}}} - d^2\right), & x > \frac{d^{\alpha_{Nr}}}{K_r P_r} \end{cases}$$

Proof. The derivation is similar to that in [16]. □

4.2 Characterization Of Mean Rewards

We will discuss how the arms of MAB framework is characterized and their mean rewards. Arm1 represents the sub-6GHz transmission and arm2 corresponds to the mm-Wave transmission. We assume that the link, AP to user, transition from LOS to NLOS state at unknown time steps, e.g., communication link blockage due a vehicle or if user is mobile, then blockage due to a building. We assume the rewards of the arms to be the SINR coverage probability[1] of the user for a threshold γ. The mean

[1]The SINR coverege probability is defined as the probability that a typical user receives the SINR greater than a threshold. Ergodically it represents the percentage of users in the network under coverage.

of both the arms follows two-state Markov model, based on the visibility sate, with unknown transition probabilities. For arm $I_t \in \kappa$, the rewards change in the manner: $\{\mu_{r,L} \rightarrow \mu_{r,N} \rightarrow \mu_{r,L} \rightarrow ....\}$, corresponding to RAT r. Note that, transitions for both the arms occurs at the same time, as visibility state is same for both the bands.

Lemma 3. *For a user served in the sub-6GHz band from an AP at a distance x, being in visibility state v, the SINR coverage probability is given by:*

$$\mu_{s,v} = \mathbb{P}_{Cvs}(\gamma) = exp\left(-\gamma \cdot \sigma_{N,s}^2 \cdot x - \sum_{v'} A_{v'}(\gamma, x)\right),$$

where,

$$A_{v'} = \int_x^\infty \frac{\gamma x}{y + \gamma x} \Lambda'_{v's}(dy), \quad \forall \quad v' \in \{L, N\}$$

Proof. The proof is similar to that in [13]. □

Lemma 4. *For a user served in the mm-Wave band from an AP at a distance x, being in visibility state v, the SINR coverage probability is given by:*

$$\mu_{m,v} = \mathbb{P}_{Cvm}(\gamma) = exp\left(-\frac{\gamma \cdot x \cdot \sigma_{N,M}^2}{G_0} - B_1(\gamma, x) - B_2(\gamma, x)\right),$$

$$with \quad B_1(\gamma, x) = \sum_{k=1}^{4} \left(-b_k \int_x^\infty \frac{\alpha_k \gamma x}{y + \alpha_k \gamma x} \Lambda'_{vm}(dy)\right),$$

$$and \quad B_2(\gamma, x) = \sum_{k=1}^{4} \left(-b_k \int_x^\infty \frac{\alpha_k \gamma x}{y + \alpha_k \gamma x} \Lambda'_{v'm}(dy)\right)$$

Proof. The proof is similar to that in [13]. □

Chapter 5

Simulation Results

5.1 RAT Selection In Wireless Networks

To assess the efficacy of TS-CD algorithm, we compare it with algorithms such as DTS, D-UCB, Rexp3 and PHT-UCB and also with a static association strategy (Fixed association) based on the maximum received power in the respective bands [13]. We observe the system for T = $1e4$ time-steps, with the user making a choice at every step. Further we assume that radio environment is changing visibility with $\lambda_A = 1e - 5s^{-1}$. In "Fig. 5.1" we plot the time-averaged regret for different algorithms. Specifically, we see that the static association strategy(Fixed-association), which always sticks to a particular band (sub-6GHz or mm-Wave) and Rexp3 performs worst among all other strategies. Further we see that the TS-CD with $\gamma = 0$ outperforms all other bandit algorithms by achieving near optimal regret $(\rightarrow 0)$. The figure also shows the points where change has occurred and the points where the change is detected by the Page Hinkely test (PHT) algorithm as well as goodness of fit test(GOF) algorithm. The zoomed figure shows number of samples to detect the change taken by GOF test based algorithm are less than the PHT algorithm. The shaded region in the figure represents the variance. We can see that the variance of the TS-CD and PHT-UCB is much lower then the other algorithms, which means they are more robust than the passively adaptive algorithms.

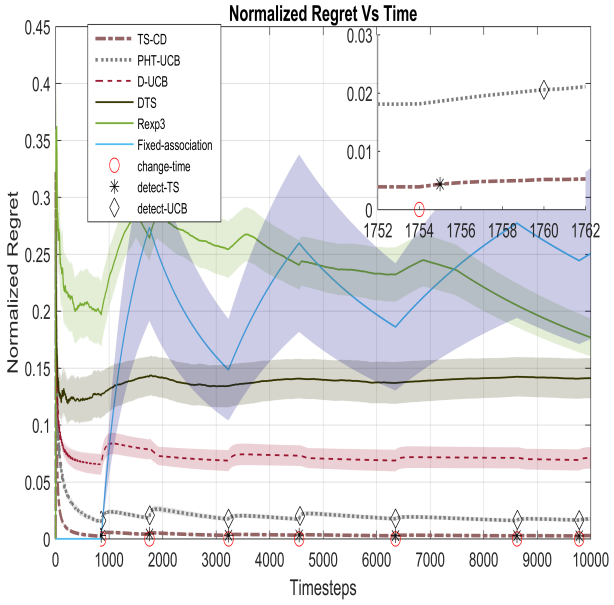

Figure 5.1: Time averaged regret for different algorithms.

This indicates that the proposed TS-CD algorithm efficiently tracks all the changes in the dynamic environment with proper hyper-parameter tuning.

5.2 Multi-Connectivity

Another application where we will test the efficacy of the proposed algorithm is multi-connectivity. First we will define a performance metric. we will define average effective throughput (AET) as a performance metric with n connections as:

$$AET(n, \gamma, r_0) = \beta \mathcal{P}_{Cn}(\gamma) + (1 - \beta)\mathcal{P}_{Rn}(r_0)$$

where, $\mathcal{P}_{Cn}(\tau) = \mathbb{P}(SINR_n > \tau)$ is SINR coverage probability defined as a typical UE receives a SINR over a given threshold τ. $\mathcal{P}_{Rn}$ is the rate coverage probability. γ is SINR threshold, n is number of connections, and β is a parameter that decides whether operator prioritizes the signal coverage or the per user throughput. To gain more insight one can refer to [24][1], we have also characterized the mean rewards from the results of the same. We assume the traffic density varies as shown in [22]. In "Fig. 5.2" we plot for $N = 2$ connections with varying UE density and fixed β, and the optimal number of connections is 2. We see that the 2 connections case coincides with the BEST strategy. Consequently, it is evident from the plot, that the 2 connection case performs slightly better than the proposed algorithm TS-CD. Also we can see that TS-CD performs better than a simple received signal-strength indicator (RSSI) based 1 connection scheme, and it is clear that, with time TS-CD learns the reward distribution and approaches the BEST and 2 connection schemes. Thus, we can infer that, under dynamic traffic intensity which does not change the optimal arm, the proposed algorithm approaches the optimal arm.

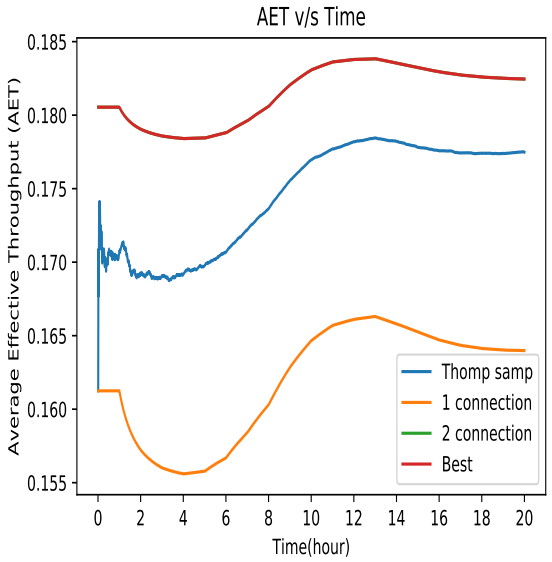

Figure 5.2: AET Performance of the proposed algorithm as compared to static association schemes with Fixed β.

Now, another type of scenario is that, when the optimal arm changes with time, such that when the rate and coverage demands of UE changes with time or the operator decides to prioritize a given type of service. In a, we plot for varying β. We can

[1]This work is under revision

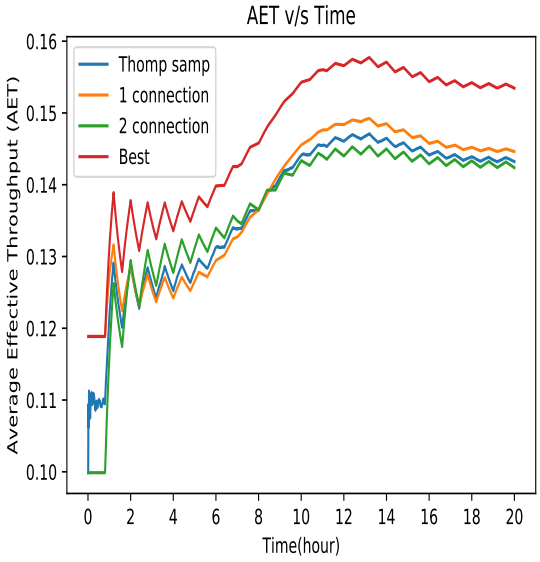

Figure 5.3: AET Performance of the proposed algorithm as compared to static association schemes with varying β.

see that initially the 2 connection case performs better, then at t = 9h, 1 connection case is performing better. The proposed algorithm TS-CD successfully tracks the change and selects the better arm.

Bibliography

[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 meta-bandits." (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 inter-cell interference coordination and intra-cell diversity." IEEE Transactions on Wireless Communications 13.12 (2014): 6655-6669.

[17] White Paper, "5G Channel Model for bands up to 100 GHz," http://www.5gworkshops.com/5gcm.html.

[18] Agrawal, Shipra, and Navin Goyal. "Analysis of thompson sampling for the multi-armed bandit problem." Conference on learning theory. 2012.

[19] Engmann, Sonja, and Denis Cousineau. "Comparing distributions: the two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test." Journal of applied quantitative methods 6.3 (2011): 1-17.

[20] Gupta, Abhishek K., Jeffrey G. Andrews, and Robert W. Heath. "Macrodiversity in cellular networks with random blockages." IEEE Transactions on Wireless Communications 17.2 (2017): 996-1010.

[21] Sun, Wen, and Jiajia Liu. "A stochastic geometry analysis of CoMP-based uplink in ultra-dense cellular networks." 2018 IEEE International Conference on Communications (ICC). IEEE, 2018.

[22] Zhou, Xinyang, and Lijun Chen. "Demand shaping in cellular networks." IEEE Transactions on Control of Network Systems 6.1 (2018): 363-374.

[23] G.Ghatak, "A change-detection based Thompson sampling framework for non-stationary bandits," submitted to IEEE Transactions on computers.

[24] G.Ghatak, Y.Sharma, K.Zaid, A.Rahman, " Elastic multi-connectivity in 5G networks." submitted to Elsevier Physical Communications.

[25] Panchenko.D (2006),Statistics for Applications[lecture notes].

[26] Szepesv´ari.C, Lattimore.T (2019), Bandit Algorithms.

[27] Gentile, Claudio, et al. "On context-dependent clustering of bandits." International Conference on Machine Learning. 2017.

[28] Richard S. Sutton, Andrew G. Barto, Reinforcement learning, an introduction.

[29] William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery, Numerical Recipes, The Art Of Scientific Computing.