A Change-Detection Based Thompson Sampling Framework for Non-Stationary Bandits

Gourab Ghatak, Member, IEEE

Abstract—We consider a non-stationary two-armed bandit framework and propose a change-detection based Thompson sampling (TS) algorithm, named TS with change-detection (TS-CD), to keep track of the dynamic environment. The non-stationarity is modeled using a Poisson arrival process, which changes the mean of the rewards on each arrival. The proposed strategy compares the empirical mean of the recent rewards of an arm with the estimate of the mean of the rewards from its history. It detects a change when the empirical mean deviates from the mean estimate by a value larger than a threshold. Then, we characterize the lower bound on the duration of the time-window for which the bandit framework must remain stationary for TS-CD to successfully detect a change when it occurs. Consequently, our results highlight an upper bound on the parameter for the Poisson arrival process, for which the TS-CD achieves asymptotic regret optimality with high probability. Finally, we validate the efficacy of TS-CD by testing it for edge-control of radio access technique (RAT)-selection in a wireless network. Our results show that TS-CD not only outperforms the classical max-power RAT selection strategy but also other actively adaptive and passively adaptive bandit algorithms that are designed for non-stationary environments.

1 INTRODUCTION

The multi-armed bandit (MAB) problem is a sequential decision-making framework where an agent chooses one or multiple actions, from a set of actions, based on the feedback of rewards from the previous choices. The MAB framework has found applications in the field of randomized clinical-trials [1], online recommendation systems [2], and wireless communications [4]. In the basic MAB setting, a decision maker plays (or chooses) one of the K independent arms (choices), and obtains a corresponding reward. The player repeats this experiment in a series of time-slots. Each of the K-independent arms is characterized by a reward distribution, which are not necessarily identical, and are unknown to the player. The goal of any bandit algorithm is to minimize the difference between the total rewards obtained by the player and the highest expected reward. To facilitate this, the player keeps a belief or value function associated to each arm at each time-slot, as the MAB algorithm evolves, and plays an arm accordingly. This inherently has an exploration-exploitation dilemma associated to it, i.e., the player has to make a decision on whether to play the arm which currently has the highest belief/value, or explore other arms to update their belief/value functions.

Related Work: In case the reward distributions of the arms remain stationary, several bandit algorithms, e.g., upper confidence bound (UCB) [5], have been proven to perform optimally. Recently, the Thompson Sampling (TS) algorithm, which was first introduced in [6], has gotten considerable interest. Several studies have shown the efficiency of the TS algorithm, albeit empirically, e.g., see [7]. In particular, for applications such as display advertising and online article recommendation, it outperforms other classical algorithms [8]. However, mathematical characterization of the TS algorithm is in general difficult due to its random-ized nature, unlike UCB, which facilitates exploration by an additional term. In this regard, Agarwal and Goyal [9] have recently derived bounds on the regret of the TS algorithm.

On the contrary, in case the reward distributions are non-stationary, bandit algorithms lose mathematical tractability. To keep track of the changing environments, in literature, two approaches have been proposed: i) passively adaptive, and ii) actively adaptive. Passively adaptive policies remain oblivious towards the time of changes and give more weights to the recent rewards with an aim to decrease the effect of the changes. Garivier and Moulines [10] have considered a scenario where the distribution of the rewards remain constant over epochs and change at unknown time instants. They have analyzed the theoretical upper and lower bounds of regret for the discounted UCB (D-UCB) and sliding window UCB (SW-UCB). Gupta et al. [11], extending the idea to Bayesian methods, have proposed dynamic TS. By assuming a Bernoulli bandit environment where the success probability evolves as a Brownian motion, the authors suggest to decay the effect of past observations in the posterior distribution of the arm being updated. This is done by applying an exponential filtering to the past observations. Raj and Kalyani [12] have formulated a Bayesian bandit algorithm for non-stationary environments. Derived from the classical TS algorithm, their proposed method - discounted TS (dTS) - works by discounting the effect of past observations. Besbes et al. [13] have developed a near optimal policy, REXP3. The authors have established lower bounds on the performance of any non-anticipating policy for a general class of non-stationary reward distributions. In principle, the REXP3 algorithms runs the classical EXP3 algorithm in different blocks by resetting the parameters in the beginning of each block. The optimal exploration exploitation tradeoff for non-stationary environments were studied by the authors in [14].

On the contrary, actively adaptive bandit algorithms track the dynamic environment and take certain actions (e.g., restarting the algorithm) when the variation of the en-
viren. Hartland et al. [15] have considered
dynamic bandits with abrupt changes in the reward gener-
ation process, and proposed an algorithm called Adapt-EvE.
It uses a Page-Hinkley statistical test (PHT) based change
point detection technique and utilizes a meta bandit formu-
lation for exploration-exploitation dilemma. However, theydid not provide a theoretical study about these procedures,
and their performance evaluation is mainly empirical. PHT
has also been used to adapt the window length of SW-
UCL [16]. However, the regret bounds of Adapt-EvE and
adaptive SW-UCL are still open problems. Two works which
are the closest to ours are the ones by Yu et al. [17], Cao et
al. [18], where the authors sense a change in the environment
by detecting a change in the empirical means of the rewards
of the arms. However, in both these works, the authors
assume a fixed number of changes with an interval to
derive the respective bounds. We make no such stringent
restriction; on the contrary, we assume random number of
changes, and statistically characterize the probability that
the changes occur more frequently than what the algorithm
can keep track of. Moreover, we provide tunable parameters
of false-alarm and missed detection probabilities in the
change detection framework, which is missing from [17],
[18]. Another strong assumption in [17] is that their algo-
rithm needs to query and observe the past rewards of some
unpicked arms, which is a very important assumption. In
fact, in most practical systems, such as the one considered
in our case-study, this may be impractical. Accordingly,
our proposal makes no such assumption. Finally, following
the change-detection framework, in contrast to [17], [18],
we employ the TS algorithm which has been empirically
shown to outperform index-based policies, but notoriously
challenging to mathematically characterize.
In summary, passively adaptive algorithms give more
tractable theoretical guarantees. However, it has been
demonstrated experimentally that actively adaptive al-
orithms outperform passively adaptive ones [19]. Follow-
ing this line of work, in this paper, we consider a two-
armed bandit setting and propose a variant of the TS algo-
rithm leveraging a mean-estimation based change-detection
framework that actively adapts the algorithm once a change
is detected. The contributions of the paper are as follows.
1.1 Contributions and Organization
• We adopt a mean-estimation based change detection
framework, where, using the reward returns of the
previous plays of the arms, the framework maintains
two sets of sequences at any time-step \( n \): i) the rewards
for the last \( n_T \) plays, termed as the test distribution
and ii) the rewards from the play \( n - n_T \) to \( n \) termed as the estimate distribution. The frame-
work detects a change when the mean of the above
two sequences differ by more than a threshold \( \Delta_C \).
Following this, we derive the bounds on \( n_T \) and \( N \) and
study the conditions under which the proposed
algorithm is able to detect a change. Accordingly, we
mathematically characterize the probability of false-
alarm and the probability of missed-detection of the
proposed algorithm.
• Based on the change-detection framework, we propose
a TS algorithm called TS with change-detection (TS-CD)
which refreshes the TS parameters when a change is
detected. To the best of our knowledge, no other work in
literature has investigated actively adaptive TS algo-
rithms, theoretically. Then, we study the probability of
the failure of the framework, by taking into account
the i) probability that changes occur faster than the
detection framework, ii) the error in estimation of the
means, and iii) the probability of missed detection.
Finally, we derive the upper bound on the regret of the
proposed TS-CD algorithm and study the probability
with which the regret follows this bound.
• To validate the efficacy of the algorithm, we employ
it in a wireless network equipped with decentralized
control for radio access technique (RAT) switching
between sub-6GHz and millimeter wave (mm-wave)
bands, where the rewards are characterized by using
stochastic geometry. We show that the proposed
framework outperforms the classical max-power band-
selection strategy and also other bandit algorithms that
are designed for tackling non-stationary environments.

2 The Two-Armed Bandit Setting

Let us consider a two-armed bandit framework, with arms
\( a_i \), where \( i \in \{1, 2\} \). At each time step \( n \), the player selects
an arm and observes a corresponding reward \( R_i(n) \). We
assume that the reward of arm \( a_i \) has a normal distribution
\( i.e. R_i(n) \sim N(\mu_i(n), \sigma^2) \), with unknown, non-stationary
mean and known, fixed variance \( \sigma^2 \). Additionally, we con-
sider a lower bound on the minimum difference between
the mean reward of the two arms at any instant of time, i.e.,

\[
|\mu_i(n) - \mu_j(n)| \geq \Delta_{\mu}, \quad i \neq j, \forall n. \tag{1}
\]

Furthermore, we assume a lower bound on the minimum
difference, \( \Delta_{\mu} \), between the mean rewards of the same arm
across the time slots when the changes occur, i.e.,

\[
|\mu_i(n) - \mu_i(k)| \in \{0\} \cup [\Delta_{\mu}, \infty), \quad n \neq k. \tag{2}
\]

In this paper, we will consider the case when \( \mu_i(n) \)s
are piece-wise stationary. The values of \( \mu_i(n) \) are assumed
to change at the unknown time-instants, \( T_{C_l} \), where \( l = 1, 2, \ldots \), with \( T_{C_0} \) assumed to be at \( n = 0 \). The change
instants follow a Poisson arrival process with parameter \( \lambda_C \),
and accordingly, the time-window between the changes, i.e.,
\( T_{C_{l+1}} - T_{C_l} \) is exponentially distributed:

\[
\mathbb{P}(T_{C_{l+1}} - T_{C_l} \leq k) = 1 - \exp(-\lambda_C k). \tag{3}
\]

The mean of both the arms are assumed to change simulta-
nously at \( T_{C_l} \).

Regret An algorithm \( \pi \) to solve the MAB problem needs
to decide at any time step \( n \), which arm to play (say \( a_{\pi(n)} \))
and obtains the reward \( R_{a_{\pi(n)}}(n) \) \( \sim N(\mu_{a_{\pi(n)}}(n), \sigma^2) \), based
on the arms chosen and the rewards obtained in time steps
\( 0, 1, 2, \ldots, n - 1 \). Let at a time-step \( n \), the \( a_i \) arm be the
optimal arm, i.e., \( \mu^*(n) = \mu_i(n) > \mu_j(n), i \neq j \). Then, the
regret of the algorithm after playing \( T \) rounds is given by:

\[
\mathcal{R}(T) = \sum_{n=0}^{T} (\mu^*(n) - \mu_{a_{\pi}(n)}(n)).
\]

1. Gaussian distribution has been previously considered in bandit
literature, e.g., [16].
The aim of MAB algorithms is to develop policies $\pi$ which bounds the regret in an expected sense, i.e., $E[R(T)]$. For that, we present the change-detection algorithm in the following section that tracks the non-stationary environment.

**TS-CD:** To actively detect the change when it occurs, we propose TS-CD in Algorithm 1, where $T_N$ is derived in (3), $n_T$ is derived in (7), and $\Delta_C$ is defined in (9). In particular, $n_T$ is the number of time-slots after a change at $T_C$, required to detect the change. Furthermore, $T_N$ refers to the minimum number of time-slots for which the MAB framework remains stationary after the detection of the change. In the next section, we will mathematically characterize $n_T$ and $T_N$ in terms of the system parameters.

**Algorithm 1 TS-CD**

1. $B_{1i} \leftarrow 0; B_{2i} \leftarrow 0; \forall i \in \{1,2,\ldots,K\}$  
   \hspace{1em} $\triangleright$ initializing parameters
2. $t \leftarrow 0$
3. while 1 do
   4. $i \leftarrow \beta(B_{2i} + 1, B_{1i} - B_{2i} + 1)$  
      \hspace{1em} $\triangleright$ draw from Beta dist.
   5. $a_j \leftarrow a_j \beta(\theta_j)$  
      \hspace{1em} $\triangleright$ choose the better arm
   6. $R_n(n) = \frac{R_{a_j}(n) - \mu_{\min} + \sigma Q^{-1}(\epsilon_b)}{\sigma}$  
      \hspace{1em} $\triangleright$ play the chosen arm evaluate the Bernoulli parameter
   7. $R^* = \text{Bern}(R_n(n))$  
      \hspace{1em} $\triangleright$ Bernoulli Trial
   8. $B_{1j} \leftarrow B_{1j} + 1 - R^*$  
      \hspace{1em} $\triangleright$ update the beta distribution
   9. $B_{2j} \leftarrow B_{2j} + R^*$  
      \hspace{1em} $\triangleright$ update the beta distribution
   10. $S_j = S_j \cup \{R_j\}$
   11. count = count + 1  
      \hspace{1em} $\triangleright$ update counter
   12. $n \leftarrow n + 1$
   13. if (count $\geq T_N$) then  
      \hspace{1em} change-detection phase starts
   14. $a_j \leftarrow \arg \max \{\mu_i(T_N)\}$  
      \hspace{1em} $\triangleright$ select the current best arm
   15. $S_j = S_j \cup \{R_j\}$  
      \hspace{1em} $\triangleright$ repeatedly play current best arm
   16. if (4) then  
      \hspace{1em} $\triangleright$ detect the change
   17. $B_{1i} \leftarrow 0; B_{2i} \leftarrow 0; i \in \{1,2\}$  
      \hspace{1em} $\triangleright$ refresh parameters
   18. count = 0  
      \hspace{1em} $\triangleright$ reset the counter
   19. end if
20. end if
21. end while

In TS-CD, first we initialize the parameters of the beta distribution for the two arms, $B_{1i}$ and $B_{2i}$, for $i \in \{1,2\}$ respectively with 0. This is consistent with the classical TS algorithm, and is due to the fact that we do not assume any prior information about the mean of the rewards. Then, at each time-step, we sample from the beta distribution for the two arms and play the arm which returns the larger sample (say arm $a_j$). Consequently, we receive a reward $R(n) = R_{a_j}(n)$. It must be noted that due to our assumption of Gaussian distributed rewards, the range of $R(n) \in (-\infty, \infty)$. However, for the sake of tractability and to employ the TS algorithm of [9], we map $R(n)$ to the range $[0, 1]$. For both the arms, let us define two boundary points $L$ and $U$ as $P(L \leq R_{a_j}(n) \leq U) \geq 1 - 2\epsilon_b$, $\forall j \in \{1,2\}$, where, $\epsilon_b$ is an arbitrary small positive number that defines the boundaries $L$ and $U$. Naturally, the values of $U$ and $L$ depend on the maximum and minimum values of $\mu_j$, respectively. Accordingly, let us define:

$$\mu_{\min} = \min \mu_j(n), \forall j, n$$

as the maximum and the minimum mean rewards of the arms across all the time instants. Thus, the values of $U$ and $L$ can be calculated as:

$$U = \sigma Q^{-1}(\epsilon_b) + \mu_{\max}$$

$$L = \mu_{\min} - \sigma Q^{-1}(\epsilon_b)$$

Accordingly, we define:

$$R_\pi(n) = \frac{R_{a_j}(n) - L}{U - L}$$

which ensures that $0 \leq R_\pi(n) \leq 1$ with a probability $1 - 2\epsilon_b$, and perform a Bernoulli trial with a success probability $R_\pi$. This step is similar to the algorithm presented in [9].

Following the result of the Bernoulli trial, the parameters of the beta distribution for the belief of the arm $a_j$ is updated, i.e., $B_{1j}$ is augmented by 1 in case the Bernoulli trial results in a failure, and the parameter $B_{2j}$ is augmented by 1 in case the the Bernoulli trial results in a success. The sequence of the rewards for each arm $j$ is stored in a reward set $S_j$. We assume that the distribution of the rewards remain stationary for $T_F \geq T_N + n_T$ time steps after every time a change occurs. After $T_N$ time-steps (as tracked by the variable count) since the last detected change, the change-detection part of the algorithm initiates (steps 13-20). For count $\geq T_N$, until the change is detected, the player repeatedly chooses the arm (say arm $a_j$) which had the best empirical mean at the time-step when count $= T_N$ (i.e., after $T_N$ time-steps since the last detected change). Note that the step 14, and consequently the step 15, is necessary since our MAB framework considered is based on TS, which is a randomized algorithm as compared to classical index-based policies. Thus, the absence of steps 14 and 15 would result in a non-zero (albeit small) probability of playing the non-optimal arm in a particular stationary regime, which would limit the change detection efficacy. The set $S_j$ is then updated in each step for $a_j$ (i.e., the optimal arm), until the change is detected. From the set $S_j$, we create two subsets: i) the test sequence: the rewards for the last $n_T$ plays of arm $a_j$, and ii) the estimate sequence: the rewards for the last $N$ rewards before the previous $n_T$ plays. In other words, in case the cardinality of $S_j$ is $L$, the test-sequence consists of $S_j(L - n_T : L)$ and the estimate sequence consists of $S_j(L - n_T - N_j : L - n_T - 1)$. The change is detected when the mean of the test sequence ($\mu_{\text{test}}$) differs from the mean of the estimate sequence ($\mu_i$) by more than $\Delta_C$. Once a change is detected, the parameters are reset (step 13).

Before proceeding to the mathematical characterization of the algorithm, we make the following important assumption on the number of plays of a TS algorithm for the stationary two-armed bandit.

**Assumption 1.** For the classical TS algorithm in the stationary two-armed bandit framework, given a set of reward distribution for the two arms (i.e., for a given $\mu_1$ and $\mu_2$), the number of plays of the two arms is approximately equal to the respective mean number of plays. In other words, across different realizations of the TS algorithm for a given $\mu_1$ and $\mu_2$, the number of plays of the arm $i$, $N_i$, follows $N_i \approx E[N_i]$. Assumption 1 is needed for Lemma[1] and we may relax...
this assumption under the guarantee that the optimal arm is played $T_N$ times, as explained later.

### 3 Mathematical Characterization of TS-CD

As mentioned in the last section, we detect a change when the empirical mean of the test sequence differs from the mean of the estimate sequence by a factor greater than $\Delta_C$. In other words, we detect a change, when for an arm $a_i$:

$$\left| \frac{1}{n_T} \sum_{p=L-n_T}^{L} S_i(p) - \frac{1}{N_i} \sum_{q=n-n_T-N_i}^{n-N_i} S_i(q) \right| \geq \Delta_C. \quad (4)$$

Here $\mu_{\text{test}}$ refers to the estimate of the test distribution and $\hat{\mu}_i$ refers to the estimate of the mean of the reward of arm $a_i$ obtained from the estimate sequence. Let us assume that a change occurs at time $T_{C_1}$ and is detected at a time $n = T_{D_1} > T_{C_1}$. For evaluating $\mu_{\text{test}}$, out of the $n_T$ elements of the test sequence, let us assume that $n_1$ samples $(X_1, X_2, \ldots, X_{n_1})$ are from the distribution $X_i \sim N(\mu_i(T_{C_1} - 1), \sigma^2)$ and $n_2$ samples $(Y_1, Y_2, \ldots, Y_{n_2})$ are from the distribution $Y_i \sim N(\mu_i(T_{C_1} + 1), \sigma^2)$. In order to simplify the notation, we denote $\mu_i(T_{C_1} - 1)$ by $\mu_x$ and $\mu_i(T_{C_1} + 1)$ by $\mu_y$. In other words, the mean of the optimal arm $a_i$ changes from $\mu_x$ to $\mu_y$ at $T_{C_1}$, with the condition that $|\mu_x - \mu_y| \geq \Delta_{\text{loc}}$. Furthermore, without loss of generality, we assume that $\hat{\mu}_i \leq \mu_x$. Thus, the detection criterion is:

$$\left| \frac{1}{n_T} \left( \sum_{p=n-n_T}^{n} X_p + \sum_{p=n-n_T+n_1}^{n} Y_p \right) - \frac{1}{N_i} \sum_{q=n-n_T-N_i}^{n-N_i} X_q \right| \geq \Delta_C. \quad (5)$$

For the mathematical characterization of the change detection framework, let us first informally introduce the following events, which we will make mathematically precise later in this section. We outline here that these events are listed in the following order since the events $E_2$ and $E_3$ become relevant if and only if $E_1$ does not occur.

- **$E_1$:** Two consecutive changes in the arms occur too often for the detection to keep track. In other words, for the TS-CD algorithm to precisely estimate the mean of the optimal arm, and then detect a change when it occurs, the MAB framework needs to be stationary for a period of time characterized by the change frequency bound, as discussed in Lemma 3.
- **$E_2$:** The estimate of the mean of arm $i$ is not accurate. That is, given that $E_1$ does not occur, and the MAB framework experiences sustained periods of stationarity before a change, the TS-CD algorithm $N_T$ number of samples to accurately estimate the mean reward of the optimal arm, as characterized in Lemma 1.
- **$E_3$:** The change detection framework is not able to detect the change. In other words, given that $E_1$ and $E_2$ do not occur, the TS-CD algorithm needs sufficient number ($n_T$) of samples to detect a change when it occurs, as characterized in Lemma 2.

Considering $E_2$, let us first define the accuracy of our estimate of $\hat{\mu}_i$.

**Definition 1.** $\hat{\mu}_i$ is said to be well-localized if for some $\epsilon \geq 0$, $\mu_i - \epsilon \leq \hat{\mu}_i \leq \mu_i + \epsilon$.

The Chernoff-Hoeffding bound states that for a large value of $N_i$, this occurs with a high probability, i.e.,

$$\mathbb{P} (\hat{\mu}_i < \mu_i + \epsilon \cup \hat{\mu}_i > \mu_i - \epsilon) \geq 1 - 2 \exp \left( -N_i \epsilon^2 \right).$$

Accordingly, in the following lemma, we will characterize the number of plays of a stationary TS algorithm so that sufficient plays of the optimal arm occurs so as to have $\hat{\mu}_i$ well-localized.

**Lemma 1.** Under Assumption 1 in a time-window where the MAB framework remains stationary, with $a_i$ being the optimal arm, the minimum number of plays of the TS-CD framework for the estimate of the mean of the best arm, $\hat{\mu}_i$ to be well-localized with probability greater than $1 - p_{\text{loc}}$ is given by:

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

where $W(\cdot)$ is the Lambert-W function.

**Proof.** The number of plays of the optimal arm, i.e., $N_i$, for the following condition to hold true: $\mathbb{P} (\hat{\mu}_i \leq \mu_x - \epsilon) < p_{\text{loc}}$, is naturally, $N_i \geq \frac{1}{\epsilon} \ln \frac{1}{p_{\text{loc}}}$. Note that we are only considering the lower limit of the mean of the optimal arm since the sub-optimal arm has a mean lower than the optimal arm. Now, we have to characterize the number of plays of a stationary TS algorithm so that $N_i$ plays of the optimal arm takes place, so as to have the estimate of its mean well-localized. We know that in $T$ plays of the TS algorithm, the expected number of plays of the sub-optimal arm $a_j$ is bounded as $\mathbb{E} [N_j] \leq \frac{40 \ln T}{\Delta_{\mu}^2} + \frac{48}{\Delta_{\mu}^2} + 18$. Consequently, given Assumption 1, we want the number of plays of the optimal arm to satisfy the following criterion:

$$T - \frac{40 \ln T}{\Delta_{\mu}^2} + \frac{48}{\Delta_{\mu}^2} + 18 \geq \epsilon \ln \frac{1}{p_{\text{loc}}},$$

$$\Rightarrow \exp \left( T \frac{40}{\Delta_{\mu}^2} \right) \geq \exp \left( \frac{1}{\epsilon} \ln \left( \frac{1}{p_{\text{loc}}} \right) - \frac{48}{\Delta_{\mu}^2} - 18 \right),$$

$$\Rightarrow T \geq T_N = \frac{40}{\Delta_{\mu}^2} W \left( \exp \left( \frac{40}{\Delta_{\mu}^2} \left( \frac{1}{\epsilon} \ln \left( \frac{1}{p_{\text{loc}}} \right) - \frac{48}{\Delta_{\mu}^2} \right) \right) \frac{\Delta_{\mu}^2}{40} \right).$$

**Lemma 1** characterizes the minimum number of time-steps for which, if the two-armed bandit framework remains stationary, the mean of the optimal arm is well-localized. In Fig. 1, we plot the bound on $T_N$ with varying $p_{\text{loc}}$ for different values of $\Delta_{\mu}$. Naturally, with increasing value of $p_{\text{loc}}$ or $\epsilon$, the constraint on the accuracy of the estimate of $\hat{\mu}_i$ becomes less stringent and accordingly, the lower bound on $T_N$ decreases. Similarly, the lower bound on $T_N$ also decreases with increasing $\Delta_{\mu}$. This is due to the fact that if the the mean rewards of both the arms are well-separated, the TS framework results in a larger number of plays of the first arm. This results in a better estimate of $\hat{\mu}_i$ with a lower value of $T_N$. Now, once the mean of the optimal arm
is well-localized, we have to detect a change, if it occurs. Accordingly, let us consider the case of failure of detection of change, given that \( \mu_i \) is well-localized, i.e., we consider \( E_3 \). The question that we will try to answer is: what should be the minimum number of samples \( n_T \) so as to have sufficient confidence of detection.

**Lemma 2.** Given that \( \mu_i \) is well localized at \( \mu_{x'} \) 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 in the test set is:

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

Proof.

\[
\Pr(\text{Failure}) = \Pr \left( \frac{1}{n_T} \sum_{p=n-n_T}^{n} X_p > \mu_x - \Delta_C \right), \nonumber
\]

\[
\leq \Pr \left( \frac{1}{n_T} \sum_{p=n-n_T}^{n} Y_p > \mu_x - \epsilon - \Delta_C \right), \nonumber
\]

\[
\leq \Pr \left( \frac{1}{n_T} \sum_{p=n-n_T}^{n} Y_p > \mu_y + \Delta_m - \epsilon - \Delta_C \right), \nonumber
\]

\[
\leq \exp \left( - (\Delta_m - \epsilon - \Delta_C)^2 n_T \right). \tag{8}
\]

Next we describe the choice of \( \Delta_C \). The answer to this lies in the tolerable maximum false-alarm rate. Let the acceptable probability of false alarm be \( \mathcal{P}_F \). False alarm occurs when a change is detected even though all the samples of the test distribution are from \( X \).

\[
\Pr(\text{False Alarm}) = \Pr \left( \frac{1}{n_T} \sum_{p=n-n_T}^{n} X_p > \mu_x - \Delta_m \right), \nonumber
\]

\[
\leq \Pr \left( \frac{1}{n_T} \sum_{p=n-n_T}^{n} X_p > \mu_x + \epsilon - \Delta_m \right), \nonumber
\]

\[
= Q \left( \frac{\sqrt{n_T} (\epsilon + \Delta_m)}{\sigma} \right). \nonumber
\]

Equating this to \( \mathcal{P}_F \), we get:

\[
\Delta_C = \frac{\sigma Q^{-1}(\mathcal{P}_F)}{\sqrt{n_T}} - \epsilon. \tag{9}
\]

Substituting this in (8), we have:

\[
\Pr(\text{Failure with } n_T \text{ samples}) \leq \exp \left( - (\Delta_m - \frac{\sigma Q^{-1}(\mathcal{P}_F)}{\sqrt{n_T}})^2 n_T \right). \tag{10}
\]

Now equating this to \( \mathcal{P}_M \), we get:

\[
\exp \left( - (\Delta_m - \frac{\sigma Q^{-1}(\mathcal{P}_F)}{\sqrt{n_T}})^2 n_T \right) = \mathcal{P}_M, \nonumber
\]

\[
\Rightarrow n_T = \frac{1}{\Delta_m} \left( \sqrt{\ln \frac{1}{\mathcal{P}_M}} + \sigma Q^{-1}(\mathcal{P}_F) \right). \tag{11}
\]

This completes the proof. \( \square \)

In Fig. 1G, we plot the lower bound on \( n_T \) with respect to the probability of missed detection, for different allowable false alarm probabilities, \( \Delta_m \), and \( \sigma \). Naturally, with a less stringent constraint on the probability of missed detection, the bound on \( n_T \) decreases. Interestingly, the effect of \( \Delta_m \) and \( \sigma \) on \( n_T \) is much more, as compared to the probability of missed detection and the probability of false alarm.

Now that we have characterized the lower bounds on \( n_T \) and \( T_N \), we study the bounds on \( \lambda_A \) so that the changes occur less frequently than the TS-CD algorithm is able to localize the optimal arm and track the changes. That is, we focus on the event \( E_1 \).

**Lemma 3.** To limit the probability of the frequency of change to \( p_{\text{change}} \), the bound on the value of \( \lambda_A \) is:

\[
\lambda_A \leq \frac{1}{n_T + T_N} \ln \left( \frac{1}{1 - p_{\text{change}}} \right). \tag{12}
\]

This directly follows from the exponential distribution of the inter-change distribution \( \mathcal{G} \). In Fig. 1G, we plot the bound on \( \lambda_A \) with respect to \( p_{\text{change}} \) for different values of \( p_{\text{loc}} \). Naturally, for less stringent requirement of \( p_{\text{change}} \), the upper-bound on \( \lambda_A \) is higher. On the other hand, as \( p_{\text{loc}} \) increases, the values of \( T_N \) decreases, and the bound on \( \lambda_A \) increases. As seen in Fig. 1A and Fig. 1B, the values of \( n_T \) is much smaller than \( T_N \), and consequently, the term \( T_N + n_T \) is approximately equal to \( T_N \). Thus, the effect of \( \mathcal{P}_F \) and \( \mathcal{P}_M \) on the bound of \( \lambda_A \) is less as compared to \( p_{\text{loc}} \).

Finally, given the framework developed so far, we derive the regret bound of the TS-CD algorithm for arbitrary values of \( p_{\text{loc}}, p_{\text{change}}, \) and \( \mathcal{P}_M \). We start with the result of Agrawal and Goyal [9] which states that the regret of a stationary TS algorithms for the two-armed bandit framework is bounded by the order of \( \mathcal{O}(\ln(T)) \). Based on this, we derive the following result.

**Theorem 1.** For the two-armed non-stationary bandit problem, with a probability

\[
p_{\text{tot}} = (1 - p_{\text{loc}}) (1 - p_{\text{change}}) (1 - \mathcal{P}_M), \tag{13}
\]

the TS-CD algorithm has expected regret bound:

\[
\mathbb{E}[R(T)] \leq \mathcal{O} \left( \ln(T N) \lambda_A T \frac{\Gamma \left( T N + n_T, \lambda_A T \right)}{\Gamma \left( T N + n_T \right)} \right), \tag{14}
\]

in time \( T \). Thus, the time expected regret is asymptotically bounded, i.e., \( \lim_{T \to \infty} \frac{1}{T} \mathbb{E}[R(T)] = 0 \).

Proof. In the TS-CD framework, each time a change occurs, the player follows the classical TS algorithm for \( T_N \) time-slots, thereby incurring a regret of \( \ln(T_N) \). Thereafter, under the assumption of well-localization of the mean rewards of the arms, the best-arm (during that stationary regime) is played repeatedly until the change is detected. Hence, in a period of \( T \) time-slots, the total regret is given by \( \eta_T \ln(T_N) \), where \( \eta_T \) is the number of changes within time \( T \). Let us recall that \( \eta_T \) is a Poisson distributed random variable. Furthermore, we recall our assumption that the bandit framework remains constant for at least \( T_N \) time-slots after a change and that we need \( n_T \) samples to detect the change. Hence, the maximum number of changes in a time-window \( T \) is bounded by \( \frac{T}{T_N + n_T} \). Accordingly, we bound the regret as:

\[
\mathbb{E}[R(T)] \leq \mathcal{O} \left( \mathbb{E}_{\eta_T} \left[ \eta_T \ln(T N) \cdot 1 \left( \eta_T \leq \frac{T}{T_N + n_T} \right) \right] \right),
\]
cations, considering the latency constraints and the over-
other enabling technologies, transmission in high-frequency
a tremendous increase in demand for data-rates. Among
The future wireless applications will be characterized by
ultra-reliable low-latency communications (URLLC appli-
case of massive machine-type communications (mMTC) and
ranges, especially in the mm-wave spectrum is a promis-
ment must necessarily be complemented by the existing
the first generation of mm-wave access point (AP) deploy-
sensitivity to blockages [20]. Consequently, it is evident that
RA T selection in the edge of a wireless network.
with higher number of candidate RATs, e.g., multiple sub-
6GHz bands, mm-wave and visible light communication etc. is an open problem as far as a theoretical analysis is concerned. The assumption of the dual RAT architecture is motivated by current studies, e.g., see [22] which provides an overview on such dual-RAT architectures that can be used to transmit control and data signals, respectively, at sub-6GHz and millimeter wave frequency bands. The important interplay between the mm-wave band and the sub-
6GHz band in a dual RAT architecture is also highlighted in the European project mmMagic [23] and the work by Kangas et al. [24].

It is evident that \( \lim_{T \to \infty} T^{-1} \mathbb{E} [\mathcal{R}(T)] = 0 \). In Fig. 1d we plot the value of time-averaged regret with respect to \( T \) for different values of \( p_{\text{loc}}, p_{\text{change}}, \text{and} \, p_{\text{M}} \). We see that as we change the values of the probability bounds from 0.01% to 1%, we see that the time-averaged regret bound decreases to 0 by \( 10^8 \) samples. In the next section, we employ the proposed TS-CD algorithm to facilitate dynamic

4 CASE-STUDY: RAT SELECTION IN THE EDGE

The future wireless applications will be characterized by a tremendous increase in demand for data-rates. Among other enabling technologies, transmission in high-frequency ranges, especially in the mm-wave spectrum is a promising solution. However, mm-wave transmissions suffer from several limitations, such as detrimental path-loss and high sensitivity to blockages [29]. Consequently, it is evident that the first generation of mm-wave access point (AP) deployment must necessarily be complemented by the existing cellular architecture. Additionally, due to the environment dynamics, such as human and vehicular blockages, the user equipment (UE) association to the different APs and to different available frequency bands need to be dynamic. In case of massive machine-type communications (mMTC) and ultra-reliable low-latency communications (URLLC) applications, considering the latency constraints and the overheads involved for connecting thousands of devices to the internet, such association must necessarily be decentralized.

In this section, we investigate a band-switching scheme modeled as the two-armed bandit problem, and study the efficacy of the TS-CD algorithm. In particular, we explore the performance of a blind RAT selection policy, in which, the BS instructs the UE to band switch to a different band without any need for a measurement gap [21]. First, let us discuss the system model under consideration. The case
ity. However, in this study, we assume that the transitions in both the bands occur simultaneously. This is because, the probability of a signal to be blocked mainly depends on the blockage process, which is independent of the carrier frequency, e.g., see [22] for a complete statistical analysis of the same. Due to the blockages, from the perspective of the typical user, $\phi$ can be further categorized into either LOS or NLOS processes: $\phi_L$ and $\phi_N$, respectively. The intensity of these modified processes are given by $p(x)$ 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 LOS function [25]:

$$p(x) = 1 \quad 	ext{if} \quad x \leq d \quad \text{and} \quad 0 \quad \text{if} \quad x > d.$$  

That is, any AP within a distance $d$ from the user is assumed to be in LOS, and any AP beyond a distance $d$ from the user is assumed to be in NLOS, where $d$ is the LOS ball radii [25]. To study the signal to interference plus noise ratio (SINR) performance, first, the path-loss processes are reformulated as one dimensional processes, $\phi_v^\prime = \{x_v^\prime \cdot x_i \in \phi_v, v \in \{L,N\}, r \in \{s,m\}$. The processes $\phi_v^\prime$ are non-homogeneous with intensity measures derived in [25].

Characterization of the Mean Rewards: Next, we discuss the band-switching scheme in the context of the MAB framework. In this scenario, arm $1$ represents the sub-6GHz transmission and arm $2$ corresponds to mm-wave transmission. We assume that the AP to user link transition from LOS to NLOS state at unknown time instants, e.g., due to the mobility of the users, resulting in the communication link being obstructed by buildings. For the experiments, we select $\lambda_{C} = 0.0005$ (i.e., on an average a change every 2e3 time-steps). This is consistent with stationary periods assumed in the literature, e.g., see [17], [18]. Let us recall here that each time-step consists of one play of the arm, which occurs at every sub-frame. In the flexible frame structure offered by 5G, we assume a sub-frame duration of 1 ms, which corresponds to the assumption of a stationary regime of about 2 seconds. Now, since our application caters to the outdoor pedestrian users, which generally move with slow speeds, we assume that the stationary regime considered in the paper holds. In particular, for the LOS regime, due to the massive antenna gain of the mm-wave transmissions, we have $\Delta_{m} \geq 0.3$. Additionally, since $n_{T} << T_{N}$, the effect of $\Delta_{m}$ is limited on the required stationary duration. Then, for an $\epsilon$ of 0.01, the stationary regime for the events $E_{1}$, $E_{2}$, and $E_{3}$ to not occur is less than 2 seconds, with $p_{loc} = 0.01$. This validates our assumption and the applicability of TS-CD. Let the rewards of the arms be represented by the SINR coverage probability $\mu_{t}$ of the user for a threshold $\gamma$. Consequently, the mean of each arm follows a two-state Markov model, based on the visibility state. For arm $a_{s}$, corresponding to RAT $r$, the rewards change in the manner: $\mu_{t,L} \rightarrow \mu_{t,N} \rightarrow \mu_{t,L} \rightarrow \cdots$. It must be noted that a transition in one arm coincides with the transition in the other arm, since, the visibility state is the same for both the bands.

**Lemma 4.** The SINR coverage probability, given that the user is receiving services in the sub-6GHz band from an AP at a distance $x$, being in visibility state $v$, is given by:

$$\mu_{s,v} = \mathbb{P}_{C_{vs}}(\gamma) = \exp \left(-\gamma \cdot \sigma^{2}_{N,s} \cdot x - \sum_{v'} \mathbb{A}_{v'}(\gamma, x) \right),$$

where

$$\mathbb{A}_{v'} = \frac{\gamma}{x} \frac{\sigma^{2}_{N,s}}{y + \gamma x} \mathbb{N}_{v'}(dy), \quad \forall \; v' \in \{L,N\}.$$ 

Proof. The proof is similar to that in [25]. □

**Lemma 5.** The SINR coverage probability, given that the user is receiving services in the mm-wave from an AP located at distance $x$, being in visibility state $v$, is given by:

$$\mu_{m,v} = \mathbb{P}_{C_{vm}}(\gamma) = \exp \left(-\gamma \cdot x \cdot \sigma^{2}_{N,m} - B_{1}(\gamma, x) - B_{2}(\gamma, x) \right),$$

with

$$B_{1}(\gamma, x) = \frac{\gamma}{x} \frac{\sigma^{2}_{N,m}}{y + \gamma x} \mathbb{N}_{vm}(dy),$$

and,

$$B_{2}(\gamma, x) = \frac{\gamma}{x} \frac{\sigma^{2}_{N,m}}{y + \gamma x} \mathbb{N}_{vm}(dy).$$

Proof. The proof follows in a similar way to that of Lemma 4. □

Before proceeding to the numerical results section, it is important to note that in general, the RAT-selection penalty is non-negligible and is associated with not only a receiver reconfiguration overhead, but also a non-zero switching delay, which we ignore in the experiment due to the blind switching policy, and to be consistent with the Algorithm presented in Section 2.2. Nevertheless, the RAT configuration delay and the overhead, e.g., for legacy switching schemes can be integrated into the proposed TS-CD framework by introducing a cost or penalty to switch arms. However, a complete analysis of the switching costs on the proposed algorithm, and the corresponding regret bounds will be treated in a future work.

**Simulation Setup and Results:** In Fig. 2 we compare the following 6 algorithms in the context of RAT switching: Actively adaptive algorithms: 1) PHT-UCB, which uses the Page-Hinkely statistic to detect the change [28], 2) TS-CD, which is the proposed algorithm in our work. Passively adaptive algorithms: 3) REXP3 [14], 4) Discounted TS (DTS) [12], 5) Discounted UCB (D-UCB) [10], and 6) Max-power association (called Fixed), which is a classical association rule in legacy wireless networks. We clearly see that the TS-CD algorithm not only outperforms the static association rule based on maximum received power, but also performs better than the other contending passively adaptive and actively adaptive bandit algorithms. In addition, we plot the respective variances of the different algorithms. The static association rule is shown to have the maximum variance among all the contending algorithms. This is mainly driven by the assumed variance of the rewards of a particular arm. In comparison, the passively adaptive algorithms have lower variance. However, in the initial stages, i.e., when the number of plays is small, a change in the environment results in an increase in the normalized regret of the passively adaptive algorithms.
On the contrary, the actively adaptive algorithms decrease the variance even more, establishing their higher stability as compared to the passively adaptive algorithms. Second, we show an example of the delay of change detection in the zoomed portion of the figure. Since in our case study, the difference between the SINR coverage probability of mm-wave RAT in LOS and NLOS is considerably large, we have a very high $\Delta_u$ and $\Delta_m$. This results in a significantly quick detection of the change as compared to the recently investigated PHT [23].

Currently, we are investigating more efficient change detection policies based on goodness of fit tests. Additionally, the sensitivity analysis of the algorithm with respect to the different system parameters of the example scenario concerned is a key direction of research, which we will address in a future work.

5 CONCLUSION

In this paper, we have investigated a change-detection based Thompson Sampling algorithm, named TS-CD, to keep track of the dynamic environment in the non-stationary two-armed bandit problem. We have derived the lower bound on the stationary regime time-window for TS-CD to efficiently detect the changes when they occur. Finally, we show that for given bounds on the frequency of changes, the proposed TS-CD algorithm reaches asymptotic optimality. To test the efficacy of the algorithm, we employ it in the RAT selection problem in a wireless network edge. We have shown that TS-CD not only outperforms the classical max-power band selection scheme, but also, it outperforms other bandit algorithms designed for dynamic environments.

REFERENCES

[1] 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.

[2] J. 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.

[3] J. Buccapatnam, et al., “Reward maximization under uncertainty: Leveraging side-observations on networks,” The Journal of Machine Learning Research, vol. 18, no. 1, pp. 7947–7980, 2017.

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

[5] E. Contal, et al., “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.

[6] 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.

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

[8] O. Chapelle and L. Li, “An empirical evaluation of thompson sampling,” in Advances in neural information processing systems, 2011, pp. 2249–2257.

[9] S. Agrawal and N. Goyal, “Analysis of thompson sampling for the multi-armed bandit problem,” in Conference on learning theory, 2012, pp. 98–110.

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

[11] N. Gupta et al., “Thompson sampling for dynamic multi-armed bandits,” in Machine Learning and Applications Workshops (ICMLA), 2011 10th International Conference on, vol. 1. IEEE, 2011, pp. 484–489.

[12] V. Raj and S. Kalyani, “Taming non-stationary bandits: A bayesian approach,” arXiv preprint arXiv:1707.09727, 2017.

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

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

[15] C. Hartland, et al., “Multi-armed bandit, dynamic environments and meta-bandits,” 2006.

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

[17] 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.

[18] K. Cao, et al., “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.

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

[20] White Paper, ”5G Channel Model for bands up to 100 GHz,” http://www.5gworkshops.com/5gcm.html, Dec. 6. 2015.

[21] F. B. Mismar, et al., “Deep learning predictive band switching in wireless networks,” arXiv preprint arXiv:1910.05305, 2019.

[22] D. Aziz, et al., “Architecture Approaches for 5G Millimetre Wave Access Assisted by 5G Low Band Using Multi-Connectivity,” in 2016 IEEE Globecom Workshops (GC Wshps), Dec 2016, pp. 1–6.

[23] H220-ICT-671650 mmMAGIC, “D3.1: Initial concepts on 5G architecture and integration,” Available Online at https://5g-mmagic.eu/, Mar. 2016.

[24] A. Kangas and T. Wigren, “Angle of arrival localization in lte using mimo pre-coder index feedback,” IEEE Communications Letters, vol. 17, no. 8, pp. 1584–1587, 2013.

[25] G. Ghassem and et al., “Coverage analysis and load balancing in hetnets with millimeter wave multi-rat small cells,” IEEE Transactions on Wireless Communications, vol. 17, no. 5, pp. 3154–3169, 2018.

[26] T. Bai and R. W. Heath, “Coverage and Rate Analysis for Millimeter-Wave Cellular Networks,” IEEE Trans. Wireless Commun., vol. 14, no. 2, pp. 1100–1114, 2015.

[27] T. Bai, R. Vaze, and R. W. Heath, “Analysis of Blockage Effects on Urban Cellular Networks,” IEEE Trans. Wireless Commun., vol. 13, no. 9, pp. 5070–5083, 2014.

[28] F. Liu, J. Lee, and N. Shroff, “A change-detection based framework for piecewise-stationary multi-armed bandit problem,” in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
This figure "Gourab.jpg" is available in "jpg" format from:

http://arxiv.org/ps/2009.02791v1
This figure "Gourab2.jpg" is available in "jpg" format from:

http://arxiv.org/ps/2009.02791v1