Improved Changepoint Detection for Piecewise i.i.d Bandits

Author names withheld
Unknown Affiliations

Abstract

We consider the setup of stochastic multi-armed bandits in the case when reward distributions are piecewise i.i.d. with unknown changepoints. Out of generality, we assume the reward distributions to be bounded and thus do not restrict to specific parametric exponential families. Due to the regret minimization objective, we study the change of mean, in the context when not only the change times are unknown, but also the mean before and after any change. We focus on the case when changes happen simultaneously on all arms, and in stark contrast with the existing literature, we target gap-dependent (as opposed to only gap-independent) regret bounds involving the magnitude of changes and optimality-gaps. We introduce two adaptations of UCB-strategies that employ scan-statistics in order to actively detect the changepoints, without knowing in advance the number of changepoints G. We also derive gap-independent regret bounds in the special case when both optimality and changepoint gaps are equal. The first strategy UCB-CPD does not know the time horizon T and achieves an $O(\sqrt{GT}\log T)$ regret bound, while the second strategy ImpCPD makes use of the knowledge of T to remove the $\log T$ dependency thereby closing an important gap with respect to the lower bound in this specific setting. Empirically, ImpCPD outperforms most of the passive and adaptive algorithms except the oracle-based algorithms that have access to the exact changepoints in all the considered environments.

1 Introduction

In this paper, we consider the piecewise i.i.d multi-armed bandit problem, an interesting variation of the stochastic multi-armed bandit (SMAB) setting in sequential decision making. The learning algorithm is provided with a set of decisions (or arms) which belong to the finite set $\mathcal{A}$. The learning proceeds in an iterative fashion, where in each time step t, the algorithm chooses an arm $i \in \mathcal{A}$ and receives a stochastic reward that is drawn from a distribution specific to the arm selected. There exist a finite number of changepoints G such that the reward distribution of *all* arms changes at those changepoints. The reward distributions of all arms are unknown to the learner. In this environment the learner has the goal of maximizing the cumulative reward at the end of the horizon T.

The piecewise i.i.d setting is extremely relevant to a lot of practical areas such as recommender systems, industrial manufacturing, and medical applications. In the healthcare domain, the non-stationary assumption is more realistic than i.i.d. assumption, and thus progress in this direction is important. An interesting use-case of this setting arises in drug-testing for a cure against a resistant bacteria, or virus such as AIDS. Here, the arms can be considered as various treatments while the feedback can be considered as to how the bacteria/virus reacts to the treatment administered. It is common in this setting that the behavior of the bacteria/virus changes after some time and thus its response to all the arms also change simultaneously.

Previous algorithms that have studied our model are either too conservative or require a broad set of assumptions. Passive algorithms like Discounted UCB (DUCB) (Kocsis and Szepesvári, 2006), Sliding Window UCB (SWUCB) (Garivier and Moulines, 2011) and Discounted Thompson Sampling (DTS) (Raj and Kalyani, 2017) do not actively try to locate changepoints and thus perform badly when changepoints are of large magnitude and occur frequently. The actively adaptive algorithm EXP3.R (Allesiardo et al., 2017) is an adaptive alternative to EXP3.S (Auer et al., 2002b) which was proposed for arbitrary changing environments. But EXP3.R is primarily intended for adversarial environments and thus is conservative when applied to a piecewise i.i.d. environment. The recently introduced actively adaptive algorithm CD-UCB (Liu et al., 2017) is an α-greedy policy that (at every timestep) samples uniform-randomly any arm for changepoint detection with α probability or plays the arm with highest UCB with $(1-\alpha)$ probability. This α

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

is hard to tune in experiments and come with limited theoretical guarantees. CD-UCB requires that the exploration parameter is set to $\alpha = 0$ for proving theoretical guarantees. CUSUM (Liu et al., 2017) is also an α-greedy policy that performs a two-sided CUSUM test to detect changepoints and empirically outperforms CD-UCB. CUSUM requires the knowledge of G and T for tuning α and its theoretical guarantees only hold for Bernoulli rewards for widely separated changepoints.

The main challenges in this setting are **(1)** whether the changepoints for each arm occur at the same time (global) or not, **(2)** how separated are the changepoints, **(3)** how large is the magnitude of each changepoint. We model these challenges into three assumptions stated in Assumptions 1, 2, and 3. All the previous algorithms require some combination of these assumptions (see Table 1). Moreover, in this paper we target gap-dependent bounds (bounds which depend on changepoint and optimality gaps) that are more informative and less worst case than gap-independent bounds. Hence, we look at this problem as solving two separate sub-goals of tracking the optimality gaps $\Delta_{i,g}^{opt}$ (Eq 4) between two changepoints and detecting changepoint gaps $\Delta_{i,g}^{chg}$ (Eq 2) for each changepoint $g = 1, 2, \ldots, G$ (under Assumptions 1, 2, and 3). None of the previous algorithms use this two sub-goal approach for solving this problem.

Our contributions are mainly threefold: *Algorithmic, Gap-dependent bound*, and *Gap-independent bound.*

1. Algorithmic: We propose two actively adaptive upper confidence bound (UCB) algorithms, referred to as UCB-Changepoint Detection (UCB-CPD) and Improved Changepoint Detector (ImpCPD), which try to locate the changepoints and adapt accordingly to minimize the cumulative regret. Our algorithms try to minimize the regret by quickly detecting small changepoint gaps and within two changepoints they quickly find the optimal arm. UCB-CPD is an anytime algorithm whereas ImpCPD is the first phase-based algorithm for this setting. Contrary to CD-UCB, CUSUM, and EXP3.R our algorithms divide the time into slices, and for each time slice for every arm $i \in \mathcal{A}$ they check the UCB, LCB mismatch to detect changepoints. While UCB-CPD checks for every such combination of time slices, ImpCPD only checks at certain estimated points in time horizon, and hence saves on computation time.

2. Gap-dependent bound: We analyze these algorithms theoretically and prove the gap-dependent regret upper bound that consist of both the changepoint gaps ($\Delta_{i,g}^{chg}$) as well as the optimality gaps ($\Delta_{i,g}^{opt}$) for each changepoint $g = 1, 2, \ldots, G$, $i \in \mathcal{A}$ in Theorem 1, and 2. In stark contrast, DUCB, SWUCB, and CUSUM only consider the minimum optimality gap ($\Delta_{\min}^{opt}$) over the entire horizon to prove their regret bounds, which scales very badly with a large number of changepoints G and large horizon T. The minimization of optimality regret depends on the hardness of the problem characterized by H_g^{opt} (introduced in Audibert et al. (2010), best-arm identification) as well as the minimum detectable changepoint gap $\Delta_{\epsilon_0,g}$ for each changepoint g. This is shown in Table 2 in the special case when all the gaps are same such that for all $i \in \mathcal{A}, g \in \mathcal{G}$, $\Delta_{i,g}^{opt} = \Delta_{i,g}^{chg} = \Delta_{\epsilon_0,g} = \Delta_{\min}^{opt}$ and $H_g^{opt} = \frac{K}{(\Delta_{\min}^{opt})^2}$. Such a setup when all the gaps are same and scales as $\Omega(\sqrt{1/T})$ is natural in view of existing works like Audibert and Bubeck (2009), Auer and Ortner (2010). This directly follows from the existing stochastic bandit minimax results from algorithms like MOSS (Audibert and Bubeck, 2009), UCBI (Auer and Ortner, 2010), OCUCB (Lattimore, 2016) from which we draw connections to current piecewise i.i.d algorithms like DUCB, SWUCB, EXP3.R, CUSUM. We also prove a gap-dependent and independent lower bound of independent interest for an optimal oracle policy π^* as opposed to Garivier and Moulines (2011) which proves a lower bound for a policy π only when $G = 2$ (Theorem 3). Note, that an oracle policy π^* has access to the exact changepoints, where it is restarted with its past history of interactions erased. These have not been explored by any piecewise i.i.d literature.

3. Gap-independent bound: We prove gap-independent regret upper bound of $O(\sqrt{GT}\log T)$ for UCB-CPD and an order optimal regret upper bound of $O(\sqrt{GT})$ for ImpCPD algorithm with the additional knowledge of T in the special case when both the optimality and changepoints gaps are equal as shown in Corollary 1, and 2. Hence, ImpCPD has the best theoretical guarantees amongst all the existing algorithms (see Table 2) in a particular setting. The tighter regret bound for ImpCPD follows from the proof technique of UCBI (with additional union bounds). We show that UCB-CPD and ImpCPD perform very well across diverse piecewise i.i.d environments. In all the considered environments ImpCPD outperforms most of the passively adaptive and actively adaptive algorithms except the oracle algorithms. Also, both our algorithms do not require the knowledge of the number of changepoints G for optimal performance and UCB-CPD also does not require the knowledge of horizon T, as opposed to other algorithms (see Table 1). A detailed table on gap-dependent bound can be found in Table 3, Appendix A.

Table 1: Comparison of Algorithms

Algorithm	Type	T	G	Assumptions
ImpCPD (ours)	Active	Y	N	1, 2, 3
UCB-CPD (ours)	Active	N	N	1, 2, 3
CUSUM	Active	Y	Y	1, 2, 3
EXP3.R	Active	Y	N	3
DUCB	Passive	Y	Y	1, 3
SWUCB	Passive	Y	Y	1, 3
Lower Bound	Oracle	Y	Y	1, 3

Table 2: Regret Bound of Algorithms

Algorithm	Gap-Dependent	Gap-Independent
ImpCPD (ours)	$O(\sum_{g=1}^{G} \frac{\log(T/H_g^{opt})}{(\Delta_{\min}^{opt})})$	$O\left(\sqrt{GT}\right)$
UCB-CPD (ours)	$O(\sum_{g=1}^{G} \frac{\log T}{(\Delta_{\min}^{opt})})$	$O\left(\sqrt{GT}\log T\right)$
CUSUM	$O(\frac{G\log T}{(\Delta_{\min}^{opt})})$	$O\left(\sqrt{GT\log\frac{T}{G}}\right)$
EXP3.R	N/A	$O\left(G\sqrt{T\log T}\right)$
DUCB	$O(\frac{\sqrt{GT}\log T}{(\Delta_{\min}^{opt})})$	$O\left(\sqrt{GT}\log T\right)$
SWUCB	$O(\frac{\sqrt{GT\log T}}{(\Delta_{\min}^{opt})})$	$O\left(\sqrt{GT\log T}\right)$
Lower Bound	$O(\sum_{g=1}^{G} \frac{\log(T/(GH_g^{opt}))}{\Delta_{\min}^{opt}})$	$\Omega\left(\sqrt{GT}\right)$

The rest of the paper is organized as follows. We first state the notations, definitions, and assumptions required for this setting in Section 2. Then we define our problem statement in Section 3 and in Section 4 we present the changepoint detection algorithms. Section 5 contains our main result, remarks and discussions, Section 6 contains numerical simulations where we test our proposed algorithms. We survey related works in Section 7 and finally we conclude in Section 8. The proofs are provided in Appendices.

2 Notations, Assumptions and Definitions

T denotes the time horizon. $\mathcal{A}$ denotes the finite set of arms with individual arm indexed by i such that $i = 1, \ldots, K$. We assume that the total number of arms is constant throughout the time horizon and $|\mathcal{A}| = K$. Each arm has a piecewise i.i.d distribution associated with it. The finite set $\mathcal{G}$ denotes the set of changepoints indexed by $g = 1, \ldots, G$, and t_g denote the time step when the changepoint g occurs for $g \in \mathcal{G}$. Let $t_0 = 1$ by convention, so that $t_0 < t_1 < \cdots < t_G$. Also we introduce the time interval (piece) $\rho_g = [t_g, t_{g+1} - 1]$, so the piece ρ_g starts at t_g and ends at $t_{g+1} - 1$. Note that the learner does not know these changepoints or the total number of changepoints in the environment. The reward drawn for the i-th arm, when it is pulled, for the t-th time instant is denoted by $X_{i,t}$. We assume all rewards are bounded in $[0, 1]$. $N_{i,\tau_g:\tau_{g+1}-1}$ denotes the number of times arm i has been pulled between τ_g to $\tau_{g+1} - 1$ timesteps for any sequence of increasing $(\tau_g)_g$ of integers. Also, we define $\hat{\mu}_{i,\tau_g:\tau_{g+1}-1}$ as the empirical mean of the arm i between τ_g to $\tau_{g+1} - 1$ timesteps. Let $\Delta_{\epsilon_0,g}$ be the minimum changepoint gap at t_g that can be detected with $(1 - \epsilon_0)$ probability by collecting a sample of size $n_{\epsilon_0,g}$. Note, that $n_{\epsilon_0,g}$, and $\Delta_{\epsilon_0,g}$ is not indexed by i as it is common for any arm $i \in \mathcal{A}$ satisfying the condition.

Definition 1. *We consider that on each arm i, the process generating the reward is piecewise mean constant according to the sequence $(t_g)_g$. That is, if $\mu_{i,t}$ denotes the mean reward of arm i at time t, then $\mu_{i,t}$ has same value for all $t \in \rho_g$. Therefore, we define t_g as,*

$$t_g = \min\{t > t_{g-1} : \exists i, \mu_{i,t-1} \neq \mu_{i,t}\}$$

Assumption 1. ***(Global changepoint)*** *We assume the global changepoint setting, that is $t_g = t$ implies $\mu_{i,t-1} \neq \mu_{i,t}$ for all $i \in \mathcal{A}$ and the learner does not know the $t_g, \forall g \in \mathcal{G}$.*

Definition 2. *Let the changepoint gap at t_g for an arm $i \in \mathcal{A}$ between the segments ρ_g and ρ_{g+1} be denoted as,*

$$\Delta_{i,g}^{chg} = |\mu_{i,g} - \mu_{i,g+1}|.$$

Lemma 1. ***(Minimum samples)*** *To detect a minimum changepoint of magnitude $\Delta_{\epsilon_0,g}$ between timestep t_0 and t with probability $(1 - \epsilon_0)$, it is sufficient to collect a minimum sample of $n_{\epsilon_0,g} \geq \frac{\log(\frac{2(t-t_0)^2}{\epsilon_0})}{2\left(\Delta_{\epsilon_0,g}\right)^2}$ for an arm i, such that $t_0 < t$, and $\epsilon_0 \in (0, 1)$.*

Proof 1. *The proof of Lemma 1 is given in Appendix C.*

Assumption 2. ***(Changepoints separation)*** *We assume that for every two consecutive segments ρ_g and $\rho_{g+1}, \forall g = 1, 2, \ldots, G-1$ all the three changepoints t_{g-1}, t_g and t_{g+1} satisfy the following condition,*

$$t_g + d\,(t_g - t_{g-1}) \leq t_g + \eta(t_{g+1} - t_g) = \eta t_{g+1} + (1 - \eta)t_g$$

where $\eta \in (0, 1)$, and $d(t_g - t_{g-1})$ is the maximal delay of a detection strategy in detecting a changepoint at t_g.

Definition 3. *We define a changepoint gap $\Delta_{i,g}^{chg}$ at changepoint $g \in \mathcal{G}$ for an arm $i \in \mathcal{A}$ to be an ϵ_0-optimal changepoint gap if,*

$$\Delta_{i,g}^{chg} \geq \Delta_{\epsilon_0,g}.$$

where $\Delta_{\epsilon_0,g}$ is as defined in Lemma 1.

Assumption 3. ***(Minimum gap)*** *We assume that $\forall g \in \mathcal{G}$ the changepoint gaps $\Delta_{i,g}^{chg}, \forall i \in \mathcal{A}$ are ϵ_0-optimal gaps.*

Lemma 2. ***(Detection Delay)*** *In the specific scenario, when all the changepoint gaps are same, that is $\Delta_{i,g}^{chg} = \Delta_{\epsilon_0,g} = \sqrt{\frac{K\log(T/G)}{T/G}}, \forall i \in \mathcal{A}, \forall g \in \mathcal{G}$, then to detect a changepoint of magnitude $\Delta_{\epsilon_0,g}$ with probability $(1 - \epsilon_0)$, a detection strategy starting after t_{g-1} suffers a maximum delay of,*

$$d(t_g - t_{g-1}) \leq \left(\frac{C_\eta K \log(\frac{T^2}{G^2\epsilon_0})}{2(\Delta_{\epsilon_0,g})^2}\right)$$

where, $C_\eta \leq 4\eta\log(T/G)$.

Discussion 1. *Thus, Assumption 2 makes sure that $t_g + d\,(t_g - t_{g-1})$ stays away from t_{g+1}. This ensures that when restarting the detection strategy from $t_g + d\,(t_g - t_{g-1})$, the detection of t_{g+1} will not be too much endangered. Moreover, all the gaps $\Delta^{chg}_{i,g}, \forall g \in \mathcal{G}, i \in \mathcal{A}$ are ϵ_0-optimal changepoint gaps following Assumption 3. An ϵ_0-optimal changepoint gap requires a minimum sample of $n_{\epsilon_0,g}$ to detect $\Delta_{\epsilon_0,g}$ with $(1-\epsilon_0)$ probability as shown in Lemma 1. A reasonable detection strategy tries to minimize the maximal delay $d(t_g - t_{g-1})$ in detection of a changepoint at t_g by achieving atmost a delay of $d\,(t_g - t_{g-1}) \leq \left(\frac{C_\eta K \log(\frac{T}{G\sqrt{\epsilon_0}})}{(\Delta_{\epsilon_0,g})^2}\right)$ (see Lemma 2). Our proposed algorithms have the constant in their delay of order $O(1)$ that is less than order $O(\log(T/G))$ of C_η (Discussion 7).*

Definition 4. *Let the optimality gap $\Delta^{opt}_{i,g}$ for an arm $i_{t'} \neq i^*_{t'}, \forall t' \in [t_{g-1}, t_g - 1]$ be,*

$$\Delta^{opt}_{i,g} = \mu_{i^*,g} - \mu_{i,g}.$$

We denote $\Delta^{chg}_{\max,g} = \max_{i\in\mathcal{A}} \Delta^{chg}_{i,g}$, $\Delta^{opt}_{\max,g} = \max_{i\in\mathcal{A}} \Delta^{opt}_{i,g}$, and $\Delta^{opt}_{\min} = \min_{i\in\mathcal{A},g\in\mathcal{G}} \Delta^{opt}_{i,g}$.

3 Problem Formulation

The objective of the learner is to minimize the cumulative regret till T, which is defined as follows:

$$R_T = \sum_{t'=1}^{T} \mu_{i^*_{t'}} - \sum_{t'=1}^{T} \mu_{\mathbb{I}_{t'}=i\neq i^*_{t'}}$$

where T is the horizon, $\mu_{i^*_{t'}}$ is the expected mean of the optimal arm at the t' timestep and $\mu_{\mathbb{I}'_t=i\neq i^*_{t'}}$ is the expected mean of the arm chosen by the learner at the t' timestep when it was not the optimal arm $i^*_{t'}$. Let $N_{i,g}$ is the number of times the learner has chosen arm i between t_{g-1} to t_g. The expected regret of an algorithm after T rounds can be written as,

$$\begin{aligned}
\mathbb{E}[R_T] &= \mathbb{E}\left[\sum_{t'=1}^{T} \mu_{i^*_{t'}} - \sum_{t'=1}^{T} \mu_{\mathbb{I}_{t'}=i\neq i^*_{t'}}\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\sum_{g=1}^{G}\sum_{t'=t_{g-1}}^{t_g} \mu_{i^*_{t'}} - \sum_{g=1}^{G}\sum_{t'=t_{g-1}}^{t_g} \mu_{\mathbb{I}_{t'}=i\neq i^*_{t'}}\right] \\
&\overset{(b)}{=} \sum_{i=1}^{K}\sum_{g=1}^{G} \Delta^{opt}_{i,g}\mathbb{E}[N_{i,g}]
\end{aligned}$$

where (a) is from Assumption 1, and (b) from Definition 4.

4 Algorithms

We first introduce the policy UCB-CPD in Algorithm 1 which is an adaptive algorithm based on the standard UCB1 (Auer et al., 2002a) approach. UCB-CPD pulls an arm at every timestep as like UCB1. It calls upon the Changepoint Detection (CPD) subroutine in Algorithm 2 for detecting a changepoint. Note, that like CD-UCB, and CUSUM the UCB-CPD also calls upon the changepoint detection subroutine at *every* timestep. UCB-CPD is an anytime algorithm which does not require the horizon as an input parameter or to tune its parameter δ. This is in stark contrast with CD-UCB, CUSUM, DUCB or SWUCB, that require the knowledge of G or T for optimal performance.

Algorithm 1: UCB with Changepoint Detection (UCB-CPD)

Input: $\delta > 0$;
Initialization: $t_s := 1$, $t_p := 1$.
Pull each arm once
for $t = K+1, .., T$ **do**
  Pull arm $j \in \arg\max_{i\in\mathcal{A}} \left\{\hat{\mu}_{i,t_s:t_p} + \sqrt{\frac{2\log(t)}{N_{i,t_s:t_p}}}\right\}$, observe reward $X_{j,t}$.
  Update $\hat{\mu}_{j,t_s:t_p} := \hat{\mu}_{j,t_s:t_p} + X_{j,t}$, $N_{j,t_s:t_p} := N_{j,t_s:t_p} + 1$.
  $t_p := t_p + 1$.
  if (CPD(t_s, t_p, δ)) == True **then**
    Restart: Set $\hat{\mu}_{i,t_s:t_p} := 0$, $N_{i,t_s:t_p} := 0$, $\forall i \in \mathcal{A}$, $t_s := 1$, $t_p := 1$.
    Pull each arm once.

Algorithm 2: Changepoint Detection(t_s, t_p, δ) (CPD)

Definition: $S_{i,t_s:t'} := \sqrt{\frac{\log(4t^2/\delta)}{2N_{i,t_s:t'}}}$.
for $i = 1, .., K$ **do**
  for $t' = t_s, .., t_p$ **do**
    if $\left(\hat{\mu}_{i,t_s:t'} + S_{i,t_s:t'} < \hat{\mu}_{i,t'+1:t_p} - S_{i,t'+1:t_p}\right)$ or $\left(\hat{\mu}_{i,t_s:t'} + S_{i,t_s:t'} < \hat{\mu}_{i,t'+1:t_p} - S_{i,t'+1:t_p}\right)$ **then**
      Return True

Algorithm 3: Changepoint Detection Improved(m, L_m, α, ψ) (CPDI)

Definition: $S_{i,t_s:L_m} = \sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2N_{i,t_s:L_m}}}$.
for $i = 1, .., K$ **do**
  for $m' = 1, \ldots, m$ **do**
    if $\left(\hat{\mu}_{i,t_s:L_{m'}} + S_{i,t_s:L_{m'}} < \hat{\mu}_{i,L_{m'}+1:L_m} - S_{i,L_{m'+1}:L_m}\right)$ or $\left(\hat{\mu}_{i,t_s:L_{m'}} - S_{i,t_s:L_{m'}} > \hat{\mu}_{i,L_{m'}+1:L_m} + S_{i,L_{m'+1}:L_m}\right)$ **then**
      Return True

Next, we introduce the actively adaptive Improved Changepoint Detector policy, referred to as ImpCPD in Algorithm 4. This algorithm is motivated from UCB-Improved in Auer and Ortner (2010) and is a phase-based algorithm. To detect

changepoints, ImpCPD calls upon an additional sub-routine, the Changepoint Detection Improved (CPDI) sub-routine in Algorithm 3 at the end of every phase. Unlike UCB-Improved, ImpCPD does not pull all arms equally at every timestep. Rather between two phases it pulls the arm that maximizes the upper confidence bound as like UCB1 in Auer et al. (2002a).

Algorithm 4: Improved Changepoint Detector (ImpCPD)

Input: Time horizon T, $0 < \gamma \leq 1$.
Initialization: $B_0 := \mathcal{A}$, $m := 0$, $\epsilon_0 := 1$, $\psi := \frac{T}{K^2 \log K}$, $\alpha := \frac{3}{2}$.

$$M := \left\lfloor \frac{1}{2} \log_{1+\gamma} \frac{T}{e} \right\rfloor, \quad \ell_0 := \left\lceil \frac{\log(\psi T \epsilon_0^2)}{2\epsilon_0} \right\rceil,$$
$$L_0 := K\ell_0, t_s := 1, t_p := 1.$$

Definition: $S_{i,t_s:t_p} := \sqrt{\frac{\alpha \log(\psi T \epsilon_m^2)}{2N_{i,t_s:t_p}}}$.
for $t = K+1, .., T$ **do**
  Pull arm $j \in \arg\max_{i \in \mathcal{A}} \left\{ \hat{\mu}_{i,t_s:t_p} + S_{i,t_s:t_p} \right\}$, observe reward $X_{j,t}$.
  $t := t+1, t_p := t_p + 1$.
  if $t \geq L_m$ and $m \leq M$ **then**
    if CPDI(m, L_m, α, ψ)==True **then**
      Restart: $B_m := \mathcal{A}$; $m := 0$.
      Set $N_{i,t_s:t_p} := 0, \forall i \in \mathcal{A}$.
      Set $\hat{\mu}_{i,t_s:t_p} := 0, \forall i \in \mathcal{A}$.
      $t_s := 1, t_p := 1$.
      Pull each arm once.
    else
      for $i \in \mathcal{A}$ **do**
        if $\hat{\mu}_{i,t_s:t_p} + S_{i,t_s:t_p} < \max_{j \in \mathcal{A}} \{\hat{\mu}_{j,t_s:t_p} - S_{j,t_s:t_p}\}$ **then**
          $|B_m| := |B_m| - 1$
      Reset Parameters:
      $\epsilon_{m+1} := \frac{\epsilon_m}{(1+\gamma)}$.
      $B_{m+1} := B_m$.
      $\ell_{m+1} := \left\lceil \frac{\log(\psi T \epsilon_m^2)}{2\epsilon_m} \right\rceil$.
      $L_{m+1} := t + |B_{m+1}| \ell_{m+1}$.
      $m := m+1$.

Also, ImpCPD employs pseudo arm elimination like CCB (Liu and Tsuruoka, 2016) algorithm such that a sub-optimal arm i is never actually eliminated but the active list B_m is just modified to control the phase length. This helps ImpCPD adapt quickly because this is a global changepoint scenario and for some sub-optimal arm the changepoint maybe detected very fast. Another important divergence from UCB-Improved is the exploration parameter $0 < \gamma \leq 1$ that controls how often the changepoint detection sub-routine CPDI is called. After every phase, we reduce ϵ_m by a factor of $(1+\gamma)$ instead of halving it (as like UCB-Improved) so that the number of pulls allocated for exploration to each arm is lesser than UCB-Improved. The CPDI sub-routine at the end of the $m-th$ phase scans statistics so that if there is a significant difference between the sample mean of any two slices then it raises an alarm.

Running time of two algorithms: UCB-CPD calls the changepoint detection at every timestep, and ImpCPD calls upon the sub-routine only at end of phases.Hence, for a fixed horizon T and K arms, UCB-CPD calls the changepoint detection subroutine $O(KT)$ times while ImpCPD calls the changepoint detection $O(K \log T)$ times, thereby substantially reducing the costly operation of calculating the changepoint detection statistics. By appropriately modifying the confidence interval, this reduction comes at no additional cost in the order of regret (see Discussion 6)

5 Main Results

Lemma 3. ***(Control of bad-event)*** *Let, $\mu_{i,g}$ be the expected mean of an arm i for the piece ρ_g, $N_{i,t_s:t}$ be the number of times an arm i is pulled from t_s till the t-th timestep such that $t > t_g$, then at the t-th timestep for all $\delta \in (0,1]$ it holds that,*

$$\mathbb{P}\Big\{\forall t' \in [t_s, t] \colon \big(\hat{\mu}_{i,t_s:t'} - S_{i,t_s:t'} > \hat{\mu}_{i,t'+1:t} + S_{i,t'+1:t}\big) \bigcup \big(\hat{\mu}_{i,t_s:t'} + S_{i,t_s:t'} < \hat{\mu}_{i,t'+1:t} - S_{i,t'+1:t}\big)\Big\} \leq \delta$$

where $S_{i,t_s:t'} = \sqrt{\frac{\log(\frac{4t^2}{\delta})}{2N_{i,t_s:t'}}}$.

Proof 2. ***(Outline)*** *We first define the bad event $\xi_{i,t}^{chg}$ which determines whether the changepoint is detected or not. Next, we use Chernoff-Hoeffding inequality along with a union bound to bound the probability of this bad event $\xi_{i,t}^{chg}$. The proof of Lemma 3 is given in Appendix E.*

Remark 1. *Choosing $\delta = \frac{1}{t}$, the above result of Lemma 3 can be reduced to $\mathbb{P}\{\xi_{i,t}^{chg}\} \leq \frac{4}{t}$, where the event $\xi_{i,t}^{chg}$ is the bad event. Note, that δ does not depend on the knowledge of horizon T.*

Theorem 1. ***(Gap-dependent bound of UCB-CPD)*** *The expected cumulative regret of UCB-CPD using the CPD is given by,*

$$\mathbb{E}[R_t] \leq \sum_{i=1}^{K} \sum_{g=1}^{G} \Big\{ 3 + \frac{8\log(t)}{\Delta_{i,g}^{opt}} + \frac{\pi^2}{3} + \frac{6\Delta_{i,g}^{opt} \log(t)}{(\Delta_{i,g}^{chg})^2} + \frac{6\Delta_{\max,g+1}^{opt} \log(t)}{(\Delta_{\epsilon_0,g})^2} + \frac{12\Delta_{i,g+1}^{opt} \log(t)}{(\Delta_{i,g}^{chg})^2} \Big\}.$$

Proof 3. ***(Outline)*** *We first bound the probability of the bad event of pulling the sub-optimal arms between two*

changepoints and the number of pulls required for each sub-optimal arm after which they can be discarded. Similarly, we bound the probability of the bad event of not detecting a changepoint and the minimum number of pulls required to detect a changepoint of sufficient gap (see Lemma 3). In all the cases above, we use Chernoff-Hoeffding inequality to bound the probability of the bad events. The proof of Theorem 1 is given in Appendix F.

Discussion 2. *In Theorem 1, the largest contributing factor to the gap-dependent regret of UCB-CPD is of the order* $O\big(\max\big\{\sum_{g=1}^{G}\frac{\Delta^{opt}_{\max,g+1}\log(T)}{(\Delta_{\epsilon_0,g})^2},\sum_{g=1}^{G}\frac{\log T}{\Delta^{opt}_{i,g}}\big\}\big)$*, lower than that of DUCB, SWUCB and same as CUSUM, when* $\forall i\in\mathcal{A},\forall g\in\mathcal{G}$, $\Delta^{opt}_{i,g}=\Delta^{chg}_{i,g}=\Delta_{\epsilon_0,g}=\Delta^{opt}_{\min}$ *(see Table 1). Note that CUSUM requires the knowledge of* G *and* T *to reach a similar bound but UCB-CPD does not.*

Theorem 2. ***(Gap-dependent bound of ImpCPD)*** *The expected cumulative regret of ImpCPD using CPDI is upper bounded by,*

$$\mathbb{E}[R_T]\leq\sum_{i\in\mathcal{A}'}\sum_{j=1}^{G}\Bigg[1+\Delta^{opt}_{i,g}+\frac{32C_1(\gamma)\,\Delta^{opt}_{i,g}\log(\frac{T}{K\sqrt{\log K}})}{(K\log K)^{-\frac{3}{2}}}+\frac{16\log(\frac{T(\Delta^{opt}_{i,g})^2}{K\sqrt{\log K}})}{(\Delta^{opt}_{i,g})}\Bigg]+\sum_{i\in\mathcal{A}'}\sum_{j=1}^{G}\Bigg[\frac{16\Delta^{opt}_{i,g+1}\log(\frac{T(\Delta^{chg}_{i,g})^2}{K\sqrt{\log K}})}{(\Delta^{chg}_{i,g})^2}+\Delta^{opt}_{i,g+1}\Bigg]+\sum_{i\in\mathcal{A}}\sum_{j=1}^{G}\Bigg[\Delta^{opt}_{\max,g+1}+\frac{16\Delta^{opt}_{\max,g+1}\log(\frac{T(\Delta^{chg}_{i,g})^2}{K\sqrt{\log K}})}{(\Delta_{\epsilon_0,g})^2}\Bigg]$$

where γ *is exploration parameter,* $C_1(\gamma)=\left(\frac{1+\gamma}{\gamma}\right)^4$, $\mathcal{A}'=\big\{i\in\mathcal{A}:\Delta^{opt}_{i,g}\geq\sqrt{\frac{e}{T}},\Delta^{chg}_{i,g}\geq\sqrt{\frac{e}{T}},\forall g\in\mathcal{G}\big\}$ *and* $\Delta^{opt}_{i,G+1}=0,\forall i\in\mathcal{A}$.

Proof 4. ***(Outline)*** *We divide the proof into two modules. In the first module, we bound the optimality regret of not pulling the optimal arm between two changepoints* $g-1$ *to* g *using steps* $3,4$ *and* 5*. In the second module we bound the changepoint regret incurred for not detecting the* g*-th changepoint using steps* $2,6$ *and* 7*. We use Chernoff-Hoeffding inequality to bound the probability of the bad events. We control the number of pulls of each sub-optimal arm using the definition of* ℓ_{m_i} *and exploration parameter* γ*. The proof of Theorem 2 is given in Appendix G.*

Discussion 3. *In Theorem 2, the largest contributing factor to the gap-dependent regret of ImpCPD is of the order* $O\big(\max\big\{\sum_{g=1}^{G}\frac{\Delta^{opt}_{\max,g+1}\log(\frac{T(\Delta^{chg}_{\max,g})^2}{K\sqrt{\log K}})}{(\Delta^{\epsilon_0}_{g})^2},\sum_{g=1}^{G}\frac{\log(\frac{T(\Delta^{opt}_{\max,g})^2}{K\sqrt{\log K}})}{(\Delta^{opt}_{i,g})}\big\}\big)$*, lower than that of CUSUM, DUCB, and SWUCB when* $\forall i\in\mathcal{A},\forall g\in\mathcal{G}$, $\Delta^{opt}_{i,g}=\Delta^{chg}_{i,g}=\Delta_{\epsilon_0,g}=\Delta^{opt}_{\min}$ *(see Table 1).*

Corollary 1. ***(Gap-independent bound of UCB-CPD)*** *In the specific scenario, when all the gaps are same, that is* $\Delta^{opt}_{i,g}=\Delta^{chg}_{i,g}=\Delta_{\epsilon_0,g}=\sqrt{\frac{K\log(T/G)}{T/G}},\forall i\in\mathcal{A},\forall g\in\mathcal{G}$ *and* $\delta=\frac{1}{t}$ *then the worst case gap-independent regret bound of UCB-CPD is given by,*

$$\mathbb{E}[R_T]\leq 3KG+\frac{KG\pi^2}{3}+32\sqrt{KGT}\log T.$$

Proof 5. *The proof of Corollary 1 is given in Appendix H.*

Discussion 4. *From Corollary 1, we see that the largest contributing factor to the gap-independent regret of UCB-CPD is of the order* $O\left(\sqrt{GT}\log T\right)$*, same as that of DUCB.*

Corollary 2. ***(Gap-independent bound of ImpCPD)*** *In the specific scenario, when all the gaps are same, that is* $\Delta^{opt}_{i,g}=\Delta^{chg}_{i,g}=\Delta_{\epsilon_0,g}=\sqrt{\frac{K\log(T/G)}{T/G}},\forall i\in\mathcal{A},\forall g\in\mathcal{G}$ *and setting* $\gamma=0.05$ *then the worst case gap-independent regret bound of ImpCPD is given by,*

$$\mathbb{E}[R_T]\leq 3G^{1.5}\sqrt{\frac{K^3\log(T/G)}{T}}+C_1G^{1.5}K^{4.5}(\log K)^2+48\sqrt{GKT}$$

where C_1 *is an integer constant.*

Proof 6. *The proof of Corollary 2 is given in Appendix I.*

Discussion 5. *From Corollary 2, we see that the largest contributing factor to the gap-independent regret of ImpCPD is of the order* $O\left(\sqrt{GT}\right)$*. This is lower than the regret upper bound of DUCB, SWUCB, EXP3.R and CUSUM (table 1).*

Discussion 6. *From Corollary 2, we see that smaller the value of* γ *the larger is the constant* C_1 *associated with the factor* $GK^{4.5}(\log K)^2$*. Note, that* γ *determines how frequently the CPDI is called by ImpCPD and by modifying the confidence interval we have been able to increase the probability of detecting the changepoint at the cost of additional regret that only scales with* K *and not with* T*.*

Discussion 7. *From Corollary 1 and 2 we see that UCB-CPD and ImpCPD perform better than a reasonable detection strategy as they have constant that has order* $O(1)$ *associated with their delays that is less than* C_η *of order* $O(\log(T/G))$ *(see Lemma 2).*

Theorem 3. ***(Lower Bounds for oracle policy)*** *The lower bound of an oracle policy* π^* *for a horizon* T*,* K *arms and* G *changepoints is given by,*

$$\mathbb{E}_{\pi^*}[R_T]_{Gap-dependent}\geq C_1\sum_{g=1}^{G}\sum_{i=1}^{K}\frac{\log\frac{T}{GH^{opt}_g}}{\Delta^{opt}_{i,g}},$$

$$\mathbb{E}_{\pi^*}[R_T]_{Gap-independent}\geq\frac{1}{20}\sqrt{GKT},$$

where, C_1 *is a constant and* $H^{opt}_g=\sum_{i=1}^{K}\frac{1}{(\Delta^{opt}_{i,g})^2}$ *is the optimality hardness of the problem.*

Proof 7. *The proof of Theorem 3 is given in Appendix J.*

Discussion 8. *This lower bound of the optimal oracle policy π^* from Theorem 3 holds for any number of changepoints as opposed to the lower bound proved in Garivier and Moulines (2011) which is valid only for $G = 2$. Hence, ImpCPD which has a gap-independent regret upper bound of $O(\sqrt{GT})$ reaches the lower bound of the policy π^* in an order optimal sense. Also, in the special case when all the gaps are same such that for all $i \in \mathcal{A}, g \in \mathcal{G}$, $\Delta_{i,g}^{opt} = \Delta_{\epsilon_0,g} = \Delta_{i,g}^{chg} = \Delta_{\min}^{opt}$ and $H_g^{opt} = \frac{K}{(\Delta_{\min}^{opt})^2}$, then ImpCPD with a gap-dependent bound of $O\big(\sum_{g=1}^{G} \frac{\log(T/H_g^{opt})}{(\Delta_{\min}^{opt})}\big)$ matches the gap dependent lower bound of π^* except the factor G in the log term (ignoring the $\log(\frac{1}{\sqrt{\log K}})$ term).*

6 Experiments

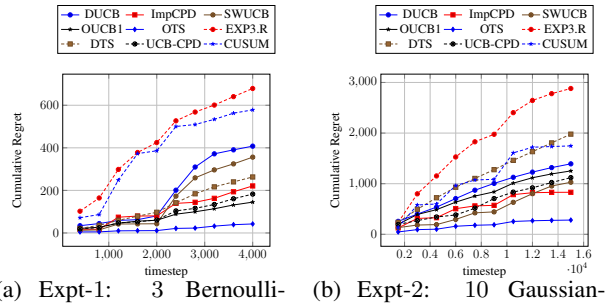
In this section, we present numerical simulations to test the performance of UCB-CPD and ImpCPD against Oracle UCB1 (OUCB1), Oracle Thompson Sampling (OTS), EXP3.R, Discounted Thompson Sampling (DTS), Discounted UCB (DUCB), Sliding Window UCB (SWUCB), and CUSUM-UCB (CUSUM) in three environments . The oracle algorithms have access to the exact changepoints and are restarted at those changepoints. For each of the experiments, we average the performance of all the algorithms over 100 independent runs. More discussions on parameter selection for the algorithms can be found in Appendix L.

Experiment 1 (Bernoulli 3 arms): This experiment is conducted to test the performance of algorithms in Bernoulli distribution over a short horizon. This experiment is conducted on a testbed of 3 arms involving Bernoulli distribution. There are 3 changepoints in this testbed where the horizon is $T = 4000$ and the expected mean of the arms changes as shown in equation 1. The experiment is shown in Figure 1(a) where we can clearly see that UCB-CPD and ImpCPD detect the changepoints at $t = 1000$ and $t = 2000$ with a small delay and restarts. However, because of the small changepoint gap at $t = 3000$ it takes some time to adapt and restart. UCB-CPD and ImpCPD perform better than all the passively adaptive algorithms like DTS, DUCB, SWUCB, and actively adaptive algorithm like EXP3.R, CUSUM and is only outperformed by OUCB1 and OTS which have access to the oracle. The performance of UCB-CPD is similar to ImpCPD in this small testbed. Because of the short horizon and a small number of arms, the adaptive algorithms CUSUM and EXP3.R are outperformed by passive algorithms DUCB, SWUCB, and DTS.

Experiment 2 (Gaussian 10 arms): This experiment is conducted to test the performance of algorithms in Gaussian distribution where the distribution flips between two

$$
\begin{aligned}
r_1 = 0.1, r_2 = 0.2, r_3 = 0.9 \quad & \textbf{if } t = [1, 1000];\\
r_1 = 0.4, r_2 = 0.9, r_3 = 0.1 \quad & \textbf{if } t = [1001, 2000];\\
r_1 = 0.5, r_2 = 0.1, r_3 = 0.2 \quad & \textbf{if } t = [2001, 3000];\\
r_1 = 0.2, r_2 = 0.2, r_3 = 0.3 \quad & \textbf{if } t = [3001, 4000]. \qquad (1)
\end{aligned}
$$

$$
\begin{aligned}
r_{1:4} = 0.4 - 0.1^{1:4}, r_5 = 0.45, r_6 = 0.55,\\
r_{7:10} = 0.6 + 0.1^{5-1:4} \textbf{ if } t = [1, 1875];\\
r_{4:1} = 0.6 + 0.1^{5-4:1}, r_5 = 0.55, r_6 = 0.45,\\
r_{7:10} = 0.4 - 0.1^{4:1} \textbf{ if } t = [1876, 5000];\\
r_{1:4} = 0.4 - 0.1^{1:4}, r_5 = 0.45, r_6 = 0.55,\\
r_{7:10} = 0.6 + 0.1^{5-1:4} \textbf{ if } t = [5001, 9000];\\
r_{4:1} = 0.6 + 0.1^{5-4:1}, r_5 = 0.55, r_6 = 0.45,\\
r_{7:10} = 0.4 - 0.1^{4:1} \textbf{ if } t = [9001, 15000]. \qquad (2)
\end{aligned}
$$



(a) Expt-1: 3 Bernoulli-distributed arms.

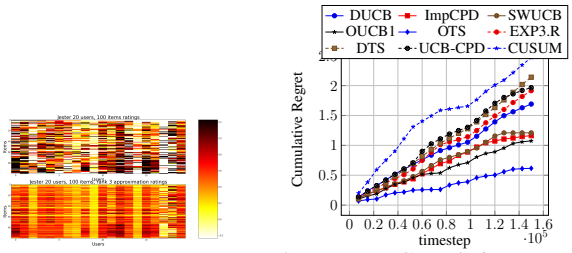
(b) Expt-2: 10 Gaussian-distributed arms.

Figure 1: Cumulative regret for various bandit algorithms on two piecewise i.i.d bandit environments.

alternating environments. This experiment involves 10 arms with Gaussian distribution. There are 3 changepoints in this testbed where the horizon is $T = 15000$ and the expected mean of the arms changes as shown in equation 2. The variance of all arms $i \in \mathcal{A}$ is set as $\sigma_i^2 = 0.25$ so that the sub-Gaussian distributions remain bounded in $[0, 1]$ with high probability. The experiment is shown in Figure 1(b) where we can clearly see that UCB-CPD and ImpCPD detect all the changepoints at $t = 1875$, $t = 5001$ and $t = 9001$ with a small delay and restarts. ImpCPD clearly performs better than all the passively adaptive algorithms like DUCB and DTS, actively adaptive algorithm like EXP3.R and is only outperformed by OTS which have access to the oracle. Also, note that ImpCPD outperforms OUCB1 (which has access to the Oracle) because of the quick detection of changepoints and usage of exploration parameters by ImpCPD. This validates the theoretical guarantees we have proved in this paper, that the regret bound of ImpCPD scales with $O(\sqrt{GT})$ while that of OUCB1 scales as $O(\sqrt{GT \log(T/G)})$ (see Proposition 1, Appendix K). In fact, when the changepoint gaps and optimality gaps are

small, the number of arms is large and the horizon is large, ImpCPD will outperform OUCB1 in all those environments.

Experiment 3 (Jester dataset): This experiment is conducted to test the performance of algorithms when our model assumptions are violated. We evaluate on the Jester dataset (Goldberg et al., 2001) which consist of over 4.1 million continuous ratings of 100 jokes from 73,421 users collected over 5 years. We evaluate on this dataset because there exist a high number of users who have rated all the jokes, and so we do not have to use any matrix completion algorithms to fill the rating matrix. The goal of the learner is to suggest the best joke when a new user comes to the system. We consider 20 users who have rated all the 100 jokes and use SVD to get a low rank approximation of this rating matrix. The rank 3 approximation is shown Figure 2(a). The dark red stripes in the figure (bottom) clearly indicates that there exist a small number of jokes which have high rating across users. Most of the users belong to three classes who prefer either joke number 49, 88, or 90. We uniform randomly sample 2 users from each of the 3 classes (49, 88, 90). Then we divide the horizon $T = 150000$ into 6 changepoints starting from $t = 1$ and at an interval of 25000 we introduce a new user from one of the three classes in round-robin fashion starting from users who prefer joke 49. We change the user at changepoints to simulate the change of distributions of arms and hence a single learning algorithm has to adapt multiple times to learn the best joke for each user. A real-life motivation of doing this may stem from the fact that running an independent bandit algorithm for each user is a costly affair and when users are coming uniform randomly a single algorithm may learn quicker across users if all the users prefer a few common items. Note, that we violate Assumption 2 (changepoints separation) and Assumption 3 (minimum gap) because the horizon is small, the number of arms is large and gaps are too small to be detectable with sufficient delay. In Figure 2(b) we see that ImpCPD outperforms most of the other algorithms except OTS and matches OUCB1 . ImpCPD is only able to detect 2 of the 6 changepoints and restart while UCB-CPD and CUSUM failed to detect any of the changepoints. Note that ImpCPD performs only slightly better than SWUCB in this testbed. This shows the importance of Assumption 2 and 3, that is when gaps are small, and changepoints are less separated, *all* the change-point detection techniques will perform badly in those regimes.



(a) Expt-3: 20 Users, 100 items, Rank 3 approximation of Jester Dataset

(b) Expt-3: Cumulative regret of different algorithms

Figure 2: Cumulative regret of algorithms in Jester dataset

7 Related Works

Existing algorithms for the piecewise i.i.d bandit setting can be broadly divided into two categories, passively adaptive and actively adaptive strategies. Passively adaptive strategies like Discounted UCB (DUCB) (Kocsis and Szepesvári, 2006), Sliding Window UCB (SWUCB) (Garivier and Moulines, 2011) and Discounted Thompson Sampling (DTS) (Raj and Kalyani, 2017) do not actively try to locate the changepoints but rather try to minimize their losses by focusing on past few observations. In Garivier and Moulines (2011) the authors showed that the regret upper bound of DUCB and SWUCB are respectively $O(\sqrt{GT}\log T)$ and $O(\sqrt{GT\log T})$, where G is the total number of changepoints and T is the time horizon which are known apriori. Furthermore, Garivier and Moulines (2011) proved that the cumulative regret in this setting for a policy π for two changepoints is lower bounded in the order of $\Omega(\sqrt{T})$. The actively adaptive strategies like Adapt-EVE (Hartland et al., 2007), Windowed-Mean Shift (Yu and Mannor, 2009), EXP3.R (Allesiardo et al., 2017), CUSUM (Liu et al., 2017) try to locate the changepoints and restart the chosen bandit algorithms. The regret bound of Adapt-EVE and DTS is still an open problem, whereas the regret upper bound of EXP3.R is $O(G\sqrt{T\log T})$ and that of CUSUM is $O(\sqrt{GT\log(T/G)})$. The regret bound of CUSUM is restricted to only Bernoulli distributions. We also implement two oracle algorithms, Oracle UCB1 (OUCB1) and Oracle Thompson Sampling (OTS) which are actively adaptive algorithms with the knowledge of the changepoints. They are restarted at the changepoints without any delay with all their past history of actions and rewards erased. OUCB1 has a regret upper bound of $O(\sqrt{GT\log(T/G)})$ (see Proposition 1, Appendix K). An extended discussion on related work can be found in Appendix B.

8 Conclusions and Future Works

We studied two UCB algorithms, UCB-CPD and ImpCPD which are adaptive and restarts once the changepoints are detected. While the anytime UCB-CPD achieves a regret of $O(\sqrt{GT}\log T)$, ImpCPD achieves the order optimal regret bound of $O(\sqrt{GT})$ which is an improvement over all the existing algorithms (in a specific setting). Furthermore, ImpCPD is faster in implementation as it is a phase-based algorithm and calls the changepoint detection sub-routine at the end of each phase while UCB-CPD calls it at every timestep and hence is slower in implementation. Empirically, ImpCPD performs very well in various environments and is only outperformed by oracle algorithms. Future works include incorporating the knowledge of localization in these adaptive algorithms or devising a tighter confidence interval by using Laplace method (Abbasi-Yadkori et al., 2011).

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pages 2312–2320.

Agrawal, R. (1995). Sample mean based index policies by $\mathcal{O}(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078.

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, pages 39.1–39.26.

Agrawal, S. and Goyal, N. (2013). Further optimal regret bounds for thompson sampling. *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, 31:99–107.

Allesiardo, R., Féraud, R., and Maillard, O. (2017). The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 3(4):267–283.

Audibert, J. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits. *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, pages 217–226.

Audibert, J., Bubeck, S., and Munos, R. (2010). Best arm identification in multi-armed bandits. *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 41–53.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2000). Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(68).

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Auer, P. and Ortner, R. (2010). Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65.

Besbes, O., Gur, Y., and Zeevi, A. J. (2014). Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. *CoRR*, abs/1405.3316.

Bubeck, S. (2010). *Bandits games and clustering foundations*. PhD thesis, University of Science and Technology of Lille-Lille I.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

Freund, Y. and Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. *European conference on computational learning theory*, pages 23–37.

Garivier, A. and Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. *Algorithmic Learning Theory - 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings*, 6925:174–188.

Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *information retrieval*, 4(2):133–151.

Hartland, C., Baskiotis, N., Gelly, S., Sebag, M., and Teytaud, O. (2007). Change point detection and meta-bandits for online learning in dynamic environments. *CAp*, pages 237–250.

Kaufmann, E., Cappé, O., and Garivier, A. (2012). On bayesian upper confidence bounds for bandit problems. *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, April 21-23, 2012*, 22:592–600.

Kocák, T., Neu, G., Valko, M., and Munos, R. (2014). Efficient learning by implicit exploration in bandit problems with side observations. *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 613–621.

Kocsis, L. and Szepesvári, C. (2006). Discounted ucb. *2nd PASCAL Challenges Workshop*, pages 784–791.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Lattimore, T. (2015). Optimally confident UCB : Improved regret for finite-armed bandits. *CoRR*, abs/1507.07880.

Lattimore, T. (2016). Regret analysis of the anytime optimally confident UCB algorithm. *CoRR*, abs/1603.08661.

Littlestone, N. and Warmuth, M. K. (1994). The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261.

Liu, F., Lee, J., and Shroff, N. B. (2017). A change-detection based framework for piecewise-stationary multi-armed bandit problem. *CoRR*, abs/1711.03539.

Liu, Y. and Tsuruoka, Y. (2016). Modification of improved upper confidence bounds for regulating exploration in monte-carlo tree search. *Theoretical Computer Science*, 644:92–105.

Maillard, O. and Munos, R. (2011). Adaptive bandits: Towards the best history-dependent strategy. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 570–578.

Mellor, J. C. and Shapiro, J. (2013). Thompson sampling in switching environments with bayesian online change point detection. *CoRR*, abs/1302.3721.

Raj, V. and Kalyani, S. (2017). Taming non-stationary bandits: A bayesian approach. *CoRR*, abs/1707.09727.

Robbins, H. (1952). Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pages 169–177. Springer.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294.

Yu, J. Y. and Mannor, S. (2009). Piecewise-stationary bandit problems with side observations. *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 1177–1184.

A Regret Bound Table

Table 3: Gap Dependent Regret Bound of Algorithms

Algorithm	Type	Gap-Dependent
ImpCPD	Active	$O\Bigg(\max \Bigg\{ \sum_{g=1}^{G} \frac{\Delta^{opt}_{\max,g+1} \log(\frac{T(\Delta^{chg}_{\max,g})^2}{K\sqrt{\log K}})}{(\Delta_{\epsilon_0,g})^2}, \sum_{g=1}^{G} \frac{\log(\frac{T(\Delta^{opt}_{\max,g})^2}{K\sqrt{\log K}})}{(\Delta^{opt}_{i,g})} \Bigg\} \Bigg)$
UCB-CPD	Active	$O\Bigg(\max \Bigg\{ \sum_{g=1}^{G} \frac{\Delta^{opt}_{\max,g+1} \log(T)}{(\Delta_{\epsilon_0,g})}, \sum_{g=1}^{G} \frac{16\log(\frac{T(\Delta^{opt}_{\min,g})^2}{K\sqrt{\log K}})}{(\Delta^{opt}_{i,g})} \Bigg\} \Bigg)$
CUSUM	Active	$O(\frac{G\log T}{(\Delta^{opt}_{\min})})$
EXP3.R	Active	N/A
DUCB	Passive	$O(\frac{\sqrt{GT}\log T}{(\Delta^{opt}_{\min})})$
SWUCB	Passive	$O(\frac{\sqrt{GT\log T}}{(\Delta^{opt}_{\min})})$
Lower Bound	Oracle	$O(\sum_{g=1}^{G} \frac{\log \frac{T}{GH^{opt}_g}}{\Delta^{opt}_{\min,g}})$

B Extended Related work

Multi-armed bandits (MAB) have been extensively studied in the *stochastic setting* where the distribution associated with each arm is fixed throughout the time horizon. Starting from the seminal works of Thompson (1933), Robbins (1952) and finally in Lai and Robbins (1985) the authors provide the asymptotic lower bound for the class of algorithms considered. The upper confidence bound (UCB) algorithms, which are a type of index-based frequentist strategy, were first proposed in Agrawal (1995) and the first finite-time analysis for the stochastic setting for this class of algorithms was proved in Auer et al. (2002a). Over the years several strong UCB type algorithms were proposed for the stochastic setting such as MOSS (Audibert and Bubeck, 2009), UCBV (Audibert et al., 2009), UCB-Improved (Auer and Ortner, 2010), OCUCB (Lattimore, 2015), etc. Further, some of the Bayesian strategies which were proposed for this setting includes the Thompson Sampling (Agrawal and Goyal, 2012), (Agrawal and Goyal, 2013) and Bayes-UCB (Kaufmann et al., 2012) algorithms.

Another setting that has greatly motivated the first studies in bandit literature is the *adversarial setting*. In this setting, at every timestep, an adversary chooses the reward for each arm and then the learner selects an arm without the knowledge of the adversary's choice. The adversary may or may not be oblivious to the learner's strategy and this forces the learner to employ a randomized algorithm to confuse the adversary. Previous works on this have focused on constructing different types of exponential weighting algorithms that are based on the Hedge algorithm that has been proposed before in Littlestone and Warmuth (1994), Freund and Schapire (1995) and analyzed in Auer et al. (2000). Further variants of this strategy called Exp3 (Auer et al., 2002b) and Exp3IX (Kocák et al., 2014) have also been proposed which incorporates different strategies for exploration to minimize the loss of the learner.

Striding between these two contrasting settings is the piecewise stochastic multi-armed bandit setting where there are a finite number of changepoints when the distribution associated with each arm changes abruptly. Hence, this setting is neither as pessimistic as adversarial setting nor as optimistic as the stochastic setting. Therefore, the two broad class of algorithms mentioned before fail to perform optimally in this setting. Several interesting solutions have been proposed before for this setting which can be broadly divided into two categories, passively adaptive and actively adaptive strategies. Passively adaptive strategies like Discounted UCB (DUCB) (Kocsis and Szepesvári, 2006), Sliding Window UCB (SWUCB) (Garivier and Moulines, 2011) and Discounted Thompson Sampling (DTS) (Raj and Kalyani, 2017) do not actively try to locate the changepoints but rather try to minimize their losses by concentrating on past few observations. In Garivier and Moulines (2011) the authors showed that the regret upper bound of DUCB and SWUCB are respectively $O\left(\sqrt{GT}\log T\right)$ and $O\left(\sqrt{GT\log T}\right)$, where G is the total number of changepoints and T is the time horizon which is known apriori. Furthermore, Garivier and Moulines (2011) showed that the cumulative regret in this setting is lower bounded in the order of $\Omega\left(\sqrt{T}\right)$. Similarly, algorithms like Restarting EXP3 (REXP3) (Besbes et al., 2014) behave pessimistically as like Exp3 but restart after pre-determined phases. Hence, REXP3 can also be termed as a passively adaptive algorithm. On the other hand,

actively adaptive strategies like Adapt-EVE (Hartland et al., 2007), Windowed-Mean Shift (Yu and Mannor, 2009), EXP3.R (Allesiardo et al., 2017), CUSUM (Liu et al., 2017) try to locate the changepoints and restart the chosen bandit algorithms. The regret bound of Adapt-EVE is still an open problem, whereas the regret upper bound of EXP3.R is $O\left(G\sqrt{T\log T}\right)$ and that of CUSUM is $O\left(\sqrt{GT\log\frac{T}{G}}\right)$. Note, that the regret bound of CUSUM is only valid for Bernoulli distributions. Also, there are Bayesian strategies like Global Change-Point Thompson Sampling (GCTS) (Mellor and Shapiro, 2013) which uses Bayesian changepoint detection to locate the changepoints. However, the regret bound of GCTS is an open problem and has not been proved yet.

Furthermore, there are the oracle algorithms which are adaptive algorithms having access to the changepoints and can be treated as baseline algorithms. Hence, they are restarted at the changepoints without any delay with all their past history of actions and rewards erased. In this paper, we implement two oracle algorithms, Oracle UCB1 (OUCB1) adapted from Auer et al. (2002a) and oracle Thompson Sampling (OTS) which is adapted from the standard Thompson Sampling algorithm (Thompson, 1933). Since these are oracle algorithms and suffer no delay, we can derive their gap-independent regret upper bounds from their stochastic counterparts. Hence, OUCB1 has a regret upper bound of $\left(\sqrt{GT\log\frac{T}{G}}\right)$ (see Proposition 1, Appendix K) and OTS have the same regret upper bound of $\left(\sqrt{GT\log\frac{T}{G}}\right)$ (Agrawal and Goyal, 2012) but empirically outperforms OUCB1.

Generally, adaptive algorithms have several advantages over passive algorithms and are more flexible. In fact, for optimal performance, the passive algorithms requires the knowledge the total number of changepoints or the frequency of change in the distribution for tuning their parameters whereas the adaptive algorithms generally require no such knowledge of the number of changepoints.

C Proof of Lemma 1

Proof 1. *We prove this lemma with a simple application of Chernoff-Hoeffding bound. Let the algorithm detect changepoint g at τ_g such that $t_0 < \tau_g \leq t$.*

Step 1. (Bad Event): *Let $\xi_{i,g}$ be the event that the learner fails to detect a change larger than $\Delta_{\epsilon_0,g}$ for the i-th arm after the g-th changepoint has occurred. Hence,*

$$\xi_{i,g} = \left\{\frac{1}{n}\sum_{s=1}^{n} X_{i,s} \ < \mu_{i,g} + \Delta_{\epsilon_0,g}\right\} \bigcup \left\{\frac{1}{n}\sum_{s=1}^{n} X_{i,s} \ > \mu_{i,g} - \Delta_{\epsilon_0,g}\right\}.$$

Step 2. (Probability of bad event): *Now we bound the probability of the bad event by using Chernoff-Hoeffding bound.*

$$\mathbb{P}\{\xi_{i,g}\} = \mathbb{P}\left\{\frac{1}{n}\sum_{s=1}^{n} X_{i,s} \ < \mu_{i,g} + \Delta_{\epsilon_0,g}\right\} \bigcup \mathbb{P}\left\{\frac{1}{n}\sum_{s=1}^{n} X_{i,s} \ > \mu_{i,g} - \Delta_{\epsilon_0,g}\right\}$$

Now, summing over all $n = [t_0, \tau_g]$ and all $\tau_g = [t_0, t]$ we can show that the probability of the bad event $\xi_{i,g}$ is upper bounded as,

$$\begin{aligned}
\mathbb{P}\{\xi_{i,g}\} &\leq \sum_{\tau_g=t_0}^{t}\sum_{n=t_0}^{\tau_g} 2\exp\left(-2\left(\Delta_{\epsilon_0,g}\right)^2 n\right) \\
&\overset{(a)}{\leq} \sum_{\tau_g=t_0}^{t}\sum_{n=t_0}^{\tau_g} 2\exp\left(-2\left(\Delta_{\epsilon_0,g}\right)^2 \frac{\log(\frac{\sqrt{2}(t-t_0)}{\sqrt{\epsilon_0}})}{\left(\Delta_{\epsilon_0,g}\right)^2}\right) \\
&\leq \sum_{\tau_g=t_0}^{t}\sum_{n=t_0}^{\tau_g} 2\frac{\epsilon_0}{2(t-t_0)^2} \\
&\leq \epsilon_0
\end{aligned}$$

where, in (a) *we substitute* $n = n_{\epsilon_0,g} \geq \dfrac{\log(\frac{\sqrt{2}(t-t_0)}{\sqrt{\epsilon_0}})}{\left(\Delta_{\epsilon_0,g}\right)^2}$.

Therefore, to detect a changepoint of magnitude greater than $\Delta_{\epsilon_0,g}$ *with probability* $(1-\epsilon_0)$*, a minimum sample of* $n_{\epsilon_0,g} \geq \dfrac{\log(\frac{\sqrt{2}(t-t_0)}{\sqrt{\epsilon_0}})}{\left(\Delta_{\epsilon_0,g}\right)^2}$ *has to be collected for an arm* $i \in \mathcal{A}$*.*

D Proof of Lemma 2

Proof 2. *From Lemma 1 we know that a minimum sample of* $n_{\epsilon_0,g}$ *is required for an arm* $i \in \mathcal{A}$ *to detect a changepoint of gap* $\Delta_{\epsilon_0,g}$*. Note, that a changepoint detection strategy may pull each arm* $i = 1,\ldots,K$*, at most* $n_{\epsilon_0,g}-1$ *times before finally detecting the changepoint for the* K*-th arm. Let,* $t_g < \tau_g \leq t_{g+1}$ *be the time the strategy detects the changepoint* g *and* $t_{g-1} < \tau_{g-1} \leq t_g$ *the detection strategy detects changepoint* $g-1$*. Note that* τ_g *and* τ_{g-1} *are random variable here.*

Hence, the delay $d(t_g - t_{g-1})$ *for the detection strategy can be upper bounded as,*

$$\begin{aligned}
d(t_g - t_{g-1}) &\leq (K-1)(n_{\epsilon_0,g}-1) + n_{\epsilon_0,g} \\
&< K n_{\epsilon_0,g} = K\frac{\log(\frac{\sqrt{2}(\tau_g-\tau_{g-1})}{\sqrt{\epsilon_0}})}{\left(\Delta_{\epsilon_0,g}\right)^2} \\
&< \frac{K\log(\frac{(\tau_g-\tau_{g-1})}{\sqrt{\epsilon_0}})}{\left(\Delta_{\epsilon_0,g}\right)^2}.
\end{aligned}$$

In the specific scenario, when the horizon T *is equally divided between* G *changepoints and all the changepoint gaps are equal such that for all* $i \in \mathcal{A}$*,* $\Delta_{i,g}^{chg} = \Delta_{\epsilon_0,g} = \sqrt{\dfrac{K\log(T/G)}{T/G}}$ *then there exist a constant* C_η *such that,*

$$\frac{C_\eta K\sqrt{2}\log(\frac{(T)}{G\sqrt{\epsilon_0}})}{\left(\Delta_{\epsilon_0,g}\right)^2} < \eta(t_{g+1}-t_g)$$

Now, to detect the changepoint g *with high probability, let* $\epsilon_0 = \frac{1}{\sqrt{T}}$*. Hence we can reduce the above expression to,*

$$
\begin{aligned}
C_\eta &< \frac{\eta(T/G)\left(\Delta_{\epsilon_0,g}\right)^2}{K\log(\frac{(T)}{G\sqrt{\epsilon_0}})} \\
&< \frac{\eta(T/G)\left(\sqrt{\frac{K\log(T/G)}{T/G}}\right)^2}{K\log(\frac{T}{G\sqrt{\epsilon_0}})} \\
&< \frac{\eta\log(T/G)}{\log(\frac{T}{G}) - \frac{1}{2}\log\epsilon_0} \\
&< \frac{\eta\log(T/G)}{\log(\frac{T}{G}) + \frac{1}{4}\log T} \\
&< \frac{4\eta\log(T/G)}{4\log(\frac{T}{G}) + \log T} \\
&< 4\eta\log(T/G)
\end{aligned}
$$

E Proof of Lemma 3

Proof 3. *We start by noting that for any arm $i \in \mathcal{A}$, we detect a change-point if,*

$$\hat{\mu}_{i,t_s:t'} - S_{i,t_s:t'} > \hat{\mu}_{i,t'+1:t} + S_{i,t'+1:t}, \text{ or } \quad \hat{\mu}_{i,t_s:t'} + S_{i,t_s:t'} < \hat{\mu}_{i,t'+1:t} - S_{i,t'+1:t}. \tag{3}$$

where for any $t' \in [t_s, t]$ we define confidence interval term $S_{i,t_s:t'}$ for an arm $i \in \mathcal{A}$ at the t-th timestep as $S_{i,t_s:t'} = \sqrt{\frac{\log(\frac{4t^2}{\delta})}{2N_{i,t_s:t'}}}$ and t_s is an arbitrary restarting timestep.

Step 1.(Define Bad Event): *We define a bad event $\xi_{i,t}^{chg}$ for each arm $i \in \mathcal{A}$ as the complementary of events in Eq (3) such that,*

$$\xi_{i,t}^{chg} = \left\{\forall t' \in [t_s, t] : \left(\hat{\mu}_{i,t_s:t'} - S_{i,t_s:t'} \leq \hat{\mu}_{i,t'+1:t} + S_{i,t'+1:t}\right) \bigcup \left(\hat{\mu}_{i,t_s:t'} + S_{i,t_s:t'} \geq \hat{\mu}_{i,t'+1:t} - S_{i,t'+1:t}\right)\right\}. \tag{4}$$

Step 2.(Define stopping times): *We define a stopping time $\tau_{i,g}$ and $\tau'_{i,g}$ for a $t' \in [t_s, t]$ such that,*

$$
\begin{aligned}
\tau_{i,g} &= \min\left\{t' : \left(\hat{\mu}_{i,t_s:t'} - S_{i,t_s:t'} \leq \hat{\mu}_{i,t'+1:t} + S_{i,t'+1:t}\right)\right\}, \\
\tau'_{i,g} &= \min\left\{t' : \left(\hat{\mu}_{i,t_s:t'} + S_{i,t_s:t'} \geq \hat{\mu}_{i,t'+1:t} - S_{i,t'+1:t}\right)\right\}.
\end{aligned}
$$

Step 3.(Minimum pulls before $\tau_{i,g}$ and $\tau'_{i,g}$): *We define $N_{i,\min_D}$ as the minimum number of pulls before $\tau_{i,g}$ such that the event $|\mu_{i,g} - \mu_{i,g-1}| > 2S_{i,t_s:\tau_{i,g}}$ is true with high probability. Then, for $N_{i,t_s:\tau_{i,g}} \geq N_{i,\min_D} = \frac{2\log(\frac{4t^2}{\delta})}{(\Delta_{i,g}^{chg})^2}$ we can show that,*

$$S_{i,t_s:\tau_{i,g}} = \sqrt{\frac{\log(\frac{4t^2}{\delta})}{2N_{i,t_s:\tau_{i,g}}}} \leq \sqrt{\frac{\log(\frac{4t^2}{\delta})}{2N_{i,\min_D}}} \leq \sqrt{\frac{(\Delta_{i,g}^{chg})^2\log(\frac{4t^2}{\delta})}{4\log(\frac{4t^2}{\delta})}} \leq \frac{\Delta_{i,g}^{chg}}{2}.$$

Similar result also holds for $\tau'_{i,g}$ such that a minimum number of pulls $N_{i,\min_D}$ is required for $|\mu_{i,g} - \mu_{i,g-1}| > 2S_{i,t_s:\tau'_{i,g}}$ to be true with high probability.

Step 4.(Bounding the probability of bad event): *In this step we bound the probability of the bad event $\xi_{i,t}^{chg}$ in Eq (4). Now from Chernoff-Hoeffding inequality we know that for any constant $\epsilon > 0$,*

$$\mathbb{P}\left\{ \left| \frac{\sum_{q=1}^{n} X_{i,q}}{n} - \mu_{i,g} \right| \geq \epsilon \right\} \leq 2\exp\left(-2\epsilon^2 n\right)$$

Next, we define the event $\Omega = \{\forall q \in [N_{i,\min_D}, t] : \hat{\mu}_{i,q:t} \leq \mu_{i,g} - S_{i,q:t}\}$. Starting with the first condition of $\xi_{i,t}^{chg}$ and applying the Chernoff-Hoeffding inequality we can show that for a $\tau_{i,g} \in [t_s, t]$,

$$\mathbb{P}\{\Omega\} \leq \sum_{n=1}^{t}\sum_{q=1}^{n} \mathbb{P}\left\{ \frac{\sum_{s=1}^{q} X_{i,s}}{q} \leq \mu_{i,g} - \frac{\log(\frac{4t^2}{\delta})}{2q} \right\} \leq \sum_{n=1}^{t}\sum_{q=1}^{n} \exp\left(-2.\frac{\log(\frac{4t^2}{\delta})}{2q} q \right) \leq \sum_{n=1}^{t}\sum_{q=1}^{n} \frac{\delta}{4t^2} \leq \frac{\delta}{4}.$$

Again, defining the event $\Omega' = \{\forall q \in [N_{i,\min_D}, t] : \hat{\mu}_{i,q:t} \geq \mu_{i,g} + S_{i,q:t}\}$ and proceeding similarly as above, we can show that for a $\tau'_{i,g} \in [t_s, t]$,

$$\mathbb{P}\{\Omega'\} \leq \sum_{n=1}^{t}\sum_{q=1}^{n} \mathbb{P}\left\{ \frac{\sum_{s=1}^{q} X_{i,s}}{q} \geq \mu_{i,g} + \frac{\log(\frac{4t^2}{\delta})}{2q} \right\} \leq \sum_{n=1}^{t}\sum_{q=1}^{n} \frac{\delta}{4t^2} \leq \frac{\delta}{4}.$$

Summing the two up, the probability that the changepoint for the arm i is not detected by first condition of $\xi_{i,t}^{chg}$ for a $t' \in [t_s, t]$ is bounded by $\frac{\delta}{2}$.

Again, for the second condition we can proceed in a similar way and show that for all $t' \in [t_s, t]$, $\tau_{i,g}, \tau'_{i,g} \in [t_s, t]$ and the two events $\Omega = \{\forall q \in [N_{i,\min_D}, t] : \hat{\mu}_{i,q:t} \leq \mu_{i,g} - S_{i,q:t}\}$ and $\Omega' = \{\forall q \in [N_{i,\min_D}, t] : \hat{\mu}_{i,q:t} \geq \mu_{i,g} + S_{i,q:t}\}$,

$$\begin{aligned}\mathbb{P}\{\Omega\} \bigcup \mathbb{P}\{\Omega'\} &\leq \sum_{n=1}^{t}\sum_{q=1}^{n} \mathbb{P}\left\{ \frac{\sum_{s=1}^{q} X_{i,s}}{q} + \frac{\log(\frac{4t^2}{\delta})}{2q} \geq \frac{\sum_{s=q+1}^{t} X_{i,s}}{t-(q+1)} - \frac{\log(\frac{4t^2}{\delta})}{2(t-(q+1))} \right\} \\ &\leq \sum_{n=1}^{t}\sum_{q=1}^{n} \frac{\delta}{4t^2} + \sum_{n=1}^{t}\sum_{q=1}^{n} \frac{\delta}{4t^2} \leq \frac{\delta}{2}.\end{aligned}$$

Hence, we can bound the probability of the bad event $\xi_{i,t}^{chg}$ for an arm $i \in \mathcal{A}$ by taking the union of all such events for all $t' \in [t_s, t]$ and for all $\tau_{i,g}, \tau'_{i,g} \in [t_s, t]$ as,

$$\mathbb{P}\{\xi_{i,t}^{chg}\} \leq \delta.$$

F Proof of Regret Bound of UCB-CPD

Proof 4. ***Step 0.(Vital Assumption for proving):*** *In this proof, we assume that all the changepoints are detected by the algorithm with some delay. Since, all the gaps are significant (Assumption 3) and there is sufficient delay between two changepoints (Assumption 2), we can take this assumption for proving the result.*

Step 1.(Some notations and definitions): *Let, $N_{i,\tau_{g-1}:\tau_g}$ denote the number of times an arm $i \in \mathcal{A}$ is pulled between τ_{g-1} to τ_g-th timestep. We recall the definition of stopping times $\tau_{i,g}$, $\tau'_{i,g}$ for an arm i from Lemma 3, step 2 as,*

$$\begin{aligned}\tau_{i,g} &= \min\left\{ t' : \left(\hat{\mu}_{i,\tau_{i,g-1}:t'} - S_{i,\tau_{i,g-1}:t'} \leq \hat{\mu}_{i,t'+1:t} + S_{i,t'+1:t} \right) \right\}, \\ \tau'_{i,g} &= \min\left\{ t' : \left(\hat{\mu}_{i,\tau_{i,g-1}:t'} + S_{i,\tau_{i,g-1}:t'} \geq \hat{\mu}_{i,t'+1:t} - S_{i,t'+1:t} \right) \right\}.\end{aligned}$$

where $t' \in [\tau_{i,g-1}, t]$. We denote $N_{i,\min_D} = \dfrac{2\log(\frac{4t^2}{\delta})}{(\Delta^{chg}_{i,g})^2}$ as the minimum number of pulls required for an arm i such that the event $|\mu_{i,g} - \mu_{i,g-1}| > 2S_{i,t_s:\tau_{i,g}}$ is true with high probability. We also recall the definition of the bad event $\xi^{chg}_{i,t}$ when the changepoint is not detected as,

$$\xi^{chg}_{i,t} = \Big\{\forall t' \in [\tau_{i,g-1}, t] : \big(\hat{\mu}_{i,\tau_{g-1}:t'} - S_{i,\tau_{i,g-1}:t'} \leq \hat{\mu}_{i,t'+1:t} + S_{i,t'+1:t}\big) \bigcup \big(\hat{\mu}_{i,\tau_{i,g-1}:t'} + S_{i,\tau_{i,g-1}:t'} \geq \hat{\mu}_{i,t'+1:t} - S_{i,t'+1:t}\big)\Big\}.$$

Also, we introduce the term τ_g as the time the algorithm detects the changepoint $g \in \mathcal{G}$ and resets as opposed to the term $\tau_{i,g}$ when the algorithm detects the changepoint for the i-th arm.

Step 2.(Define an optimality stopping time): *We define an optimality stopping time $\tau^{opt}_{i,g}$ for a $\tau^{opt}_{i,g} \in [\tau_{i,g-1}, t]$such that,*

$$\tau^{opt}_{i,g} = \min\Big\{t' : \Big(\hat{\mu}_{i,\tau_{i,g-1}:t'} < \mu_{i,g-1} + S'_{i,\tau_{i,g-1}:t'}\Big) \bigcup \Big(\hat{\mu}_{i^*,\tau_{i,g-1}:t'} > \mu_{i^*,g-1} - S'_{i^*,\tau_{i,g-1}:t'}\Big)\Big\}$$

where $S'_{i,\tau_{i,g-1}:t'} = \sqrt{\dfrac{2\log(t)}{N_{i,\tau_{i,g-1}:t'}}}$. Note, that if the two events for $\tau^{opt}_{i,g}$ come true then the sub-optimal arm is no longer pulled.

Step 3.(Minimum pulls of a sub-optimal arm): *We denote $N_{i,\min_E}$ as the minimum number of pulls required for a sub-optimal arm i such that $(\mu_{i^*,g-1} - \mu_{i,g-1}) > 2S'_{i,\tau_{i,g-1}:t'}$ is satisfied with high probability. Indeed we can show that for $N_{i,\tau_{i,g-1}:t} \geq N_{i,\min_E} = \dfrac{8\log t}{(\Delta^{opt}_{i,g})^2}$,*

$$S'_{i,\tau_{i,g-1}:t} = \sqrt{\frac{2\log(t)}{N_{i,\tau_{i,g-1}:t}}} \leq \sqrt{\frac{2\log(t)}{N_{i,\min_E}}} \leq \frac{\Delta^{opt}_{i,g}}{2}$$

Step 4.(Define the optimality bad event): *We define the optimality bad event $\xi^{opt}_{i,t}$ as,*

$$\xi^{opt}_{i,t} = \Big\{\forall t' \in [\tau_{i,c_{j-1}}, t] : \hat{\mu}_{i^*,\tau_{i,c_{j-1}}:t} + S'_{i^*,\tau_{i,c_{j-1}},t} \leq \hat{\mu}_{i,\tau_{i,c_{j-1}}:t} + S'_{i,\tau_{i,c_{j-1}},t}\Big\}$$

For the above event to be true the following three events have to be true,

$$\xi^{opt}_{i,t} \subset \Big\{\forall t' \in [\tau_{i,g-1}, t] : \{\hat{\mu}_{i,\tau_{i,g-1}:t'} \geq \mu_{i,g-1} + S'_{i,\tau_{i,g-1}:t'}\} \bigcup \{\hat{\mu}_{i^*,\tau_{i,g-1}:t'} \leq \mu_{i^*,g-1} - S'_{i^*,\tau_{i,g-1}:t'}\} \bigcup \{(\mu_{i^*,g-1} - \mu_{i,g-1}) \leq 2S'_{i,\tau_{i,g-1}:t'}\}\Big\}$$

which implies that the sub-optimal arm i has been over-estimated, the optimal arm i^* has been under-estimated or there is sufficient gap between $(\mu_{i^*,g-1} - \mu_{i,g-1})$. But, from step 3, we know that for $N_{i,\tau_{i,g-1}:t'} \geq N_{\min_E}$ the third event of $\xi^{opt}_{i,t}$ is not possible with high probability.

Step 5.(Bound the number of pulls): *Let $\tau_{i,c_0} = 1$ denoting the first timestep. Now, the total number of pulls $N_{i,1:t}$ for an arm $i \in \mathcal{A}$ till t-th timestep is given by,*

$$N_{i,1:t} = \left[1 + \sum_{t'=K+1}^{t} \mathbf{1}_{\{\mathbb{I}_{t'}=i\neq i^*_{t'}\}}\right]$$

*We consider the contribution of each changepoint $g \in \mathcal{G}, \forall g = 1, \ldots, G$ to the number of pulls $N_{i,1:t}$ and using **step 0** decompose the event $\{\mathbb{I}_{t'} = i \neq i^*_{t'}\}$ into four parts,*

$$N_{i,1:t} = \sum_{j=1}^{G}\Bigg[\underbrace{1 + \sum_{t'=\tau_{i,g-1}+K+1}^{t_g} \mathbf{1}_{\{\mathbb{I}_{t'}=i\neq i^*_{t'}, N_{i,\tau_{i,g-1}:t_g} < \max\{N_{i,\min_E}, N_{i,\min_D}\}\}}}_{\textbf{part A}}$$
$$+ \underbrace{\sum_{t'=\tau_{i,g-1}+K+1}^{t_g} \mathbf{1}_{\{\mathbb{I}_{t'}=i\neq i^*_{t'}, N_{i,\tau_{i,g-1}:t_g} \geq \max\{N_{i,\min_E}, N_{i,\min_D}\}\}}}_{\textbf{part B}} + + \underbrace{\sum_{t'=t_g}^{\tau_{i,g}} \mathbf{1}_{\{\mathbb{I}_{t'}=i\neq i^*_{t'}, N_{i,t_g:\tau_{i,g}} \geq N_{i,\min_D}\}}}_{\textbf{part D}}$$
$$+ \underbrace{\sum_{t'=t_g}^{\tau_{i,g}} \mathbf{1}_{\left\{\mathbb{I}_{t'=i\neq i*_{t'}, N_{i,t_g:\tau_{i,g}} < N_{i,min_D}, \tau_{i,g}=\tau_g\right\}} + \sum_{t'=t_g}^{\tau_g} \mathbf{1}_{\left\{\mathbb{I}_{t'=i\neq i*_{t'}, N_{i,t_g:\tau_{i,g}} < N_{i,min_D}, \tau_{i,g}<\tau_g\right\}}}_{\textbf{part C}}\Bigg]$$

*where, **part A** refers to the minimum number of pulls required before either UCB-CPD discards the sub-optimal arm i or a changepoint is detected between $\tau_{i,g-1} : t_g$, **part B** refers to the number of pulls due to bad event that UCB-CPD continues to pull the sub-optimal arm i or the changepoint is not detected between $\tau_{i,g-1} : t_g$, **part C** refers to the pulls accrued due to the worst case events when the changepoint has occurred and has not been detected by UCB-CPD from $t_g : \tau_{i,g}$ and finally, **part D** refers to number of pulls for the bad event that the changepoint is not detected from $t_g : \tau_{i,g}$.*

*Now, we will consider this parts individually and see their contribution to the number of pulls. For **part A**,*

$$\underbrace{1 + \sum_{t'=\tau_{i,g-1}+K+1}^{t_g} \mathbf{1}_{\{\mathbb{I}_{t'}=i\neq i^*_{t'}, N_{i,\tau_{i,g-1}:t_g} < \max\{N_{i,\min_E}, N_{i,\min_D}\}\}}}_{\textbf{part A}} \leq 1 + N_{i,\min_E} + N_{i,\min_D}$$
$$\overset{(a)}{=} 1 + \frac{8\log t}{(\Delta^{opt}_{i,g-1})^2} + \frac{2\log(\frac{4t^2}{\delta})}{(\Delta^{chg}_{i,g})^2}$$

*where (a) is obtained from previous step $3, 4$. Similarly, for **part B** we can show that the event,*

$$\left\{\mathbb{I}_{t'} = i \neq i^*_{t'}, N_{i,\tau_{i,g-1}:t_g} \geq \max\{N_{\min_E}, N_{i,\min_D}\}\right\} \subseteq \{\xi^{opt}_{i,t_g} + \xi^{chg}_{i,t_g}\}$$

*which is obtained from step 4. Hence, we can upper bound **part B** as,*

$$\underbrace{\sum_{t'=\tau_{i,g-1}+K+1}^{t_g} \mathbf{1}_{\{\mathbb{I}_{t'}=i\neq i^*_{t'}, N_{i,\tau_{i,g-1}:t_g} \geq \max\{N_{i,\min_E}, N_{i,\min_D}\}\}}}_{\textbf{part B}} \leq \sum_{t'=\tau_{i,g-1}+K+1}^{t_g} \xi^{opt}_{i,t'} + \sum_{t'=\tau_{i,g-1}+K+1}^{t_g} \xi^{chg}_{i,t'}.$$

*Now, **part C** consist of the worst case scenario where the changepoint has occurred and has not been detected. A good learner always try to minimize the number of pulls of the sub-optimal arms that might occur in this period from $t_g : \tau_{i,g}$. Then, we can show that the event in **part C** is the subset of the union of several worst case events.*

$$\{\mathbb{I}_{t'} = i \neq i^*_{t'}, N_{i,t_g:\tau_{i,g}} < N_{i,\min_D}, \tau_{i,g} = \tau_g\} \bigcup \{\mathbb{I}_{t'} = i \neq i^*_{t'}, N_{i,t_g:\tau_{i,g}} < N_{i,\min_D}, \tau_{i,g} < \tau_g\}$$
$$\subseteq \left\{ \mathbb{I}_{t'} = i \in \arg\max_{i \in \mathcal{A}} \left\{ \arg\max_{k \in \mathcal{A}} \{\Delta^{opt}_{k,g+1}\} \bigcup \arg\min_{k \in \mathcal{A}} \{\Delta^{chg}_{k,g}\} \bigcup \arg\max_{k \in \mathcal{A}} \{\frac{\Delta^{opt}_{k,g+1}}{\Delta^{chg}_{k,g}}\} \right\} \right\}$$

*Now, to bound the number of pulls of the sub-optimal arm i due to these several worst case events, we note that for the first event of **part C**, the changepoint is detected due to arm i. Hence,*

$$\sum_{t'=t_g}^{\tau_{i,g}} \mathbf{1}_{\left\{\mathbb{I}_{t'=i \neq i^*_{t'}, N_{i,t_g:\tau_{i,g}} < N_{i,min_D}, \tau_{i,g} = \tau_g}\right\}} \leq N_{i,\min_D}$$

*Next, for the second event of **part C**, the changepoint is detected not due to arm i, but for a different arm at a later timestep such that $\tau_g > \tau_{i,g}$. Hence,*

$$\sum_{t'=t_g}^{\tau_g} \mathbf{1}_{\left\{\mathbb{I}_{t'=i \neq i^*_{t'}, N_{i,t_g:\tau_{i,g}} < N_{i,min_D}, \tau_{i,g} < \tau_g}\right\}} \leq \max_{N_{1,t_g:\tau_g}, \ldots, N_{K,t_g:\tau_g}} \left\{ \sum_{i=1}^{K} N_{i,t_g:\tau_g} : N_{i,t_g:\tau_g} < N_{i,min_D} \right\}.$$

*Finally, for **part D** we can show that for $t' \in [t_g, \tau_{i,g}]$ or $t' \in [t_g, \tau'_{i,g}]$,*

$$\left\{\mathbb{I}_{t'} = i \neq i^*_{t'}, N_{i,t_g:\tau_{i,g}} \geq N_{\min_D}\right\} \subseteq \{\xi^{chg}_{i,\tau_{i,g}}\} \bigcup \{(\mu_{i^*,g} - \mu_{i,t_g}) < 2S_{i,t_g:\tau_{i^*,g}}\}.$$

*But from step 5 we know that for $N_{i,\tau_{g-1}:\tau_g} \geq N_{i,\min_D}$ then the event $(\mu_{i^*,g} - \mu_{i,g}) < 2S_{i,t_g:\tau_{i,g}}$ is not possible. Hence, we can show that **part D** is upper bounded by,*

$$\underbrace{\sum_{t'=t_g}^{\tau_{i,g}} \mathbf{1}_{\{\mathbb{I}_{t'}=i \neq i^*_{t'}, N_{i,t_g:\tau_{i,g}} \geq N_{i,\min_D}\}}}_{\textbf{part D}} \leq N_{i,\min_D} + \sum_{t'=t_g}^{\tau_{i,g}} \xi^{chg}_{i,t'} \leq N_{i,\min_D} + \sum_{t'=\tau_{i,g-1}}^{\tau_{i,g}} \xi^{chg}_{i,t'}$$
$$= \frac{2\log(\frac{4t^2}{\delta})}{(\Delta^{chg}_{i,g})^2} + \sum_{t'=\tau_{i,g-1}}^{\tau_{i,g}} \xi^{chg}_{i,t'}.$$

***Step 6.(Bound the expected regret):** Now, taking expectation over $N_{i,1:t}$, and considering the contributions form **parts A, B, C, D** we can show that,*

$$\mathbb{E}[N_{i,1:t}] \leq \sum_{j=1}^{G} \Bigg[\underbrace{1 + N_{i,\min_E} + \sum_{t'=N_{i,\min_E}}^{t_g} \mathbb{P}\{\xi^{opt}_{i,t'}\}}_{\textbf{A1) expected pulls from } \tau_{i,g-1}:t_g \textbf{ for optimality detection}} + \underbrace{N_{i,\min_D} + \sum_{t'=N_{i,\min_D}}^{t_g} \mathbb{P}\{\xi^{chg}_{i,t'}\}}_{\textbf{B1) expected pulls from } \tau_{i,g-1}:t_g \textbf{ for changepoint detection}}$$
$$+ \underbrace{N_{i,\min_D} + \max_{N_{1,t_g:\tau_g}, \ldots, N_{K,t_g:\tau_g}} \left\{ \sum_{i=1}^{K} N_{i,t_g:\tau_g} : N_{i,t_g:\tau_g} < N_{i,min_D} \right\}}_{\textbf{C1) worst case expected pulls from } t_g:\tau_{i,g}} + \underbrace{N_{i,\min_D} + \sum_{t'=N_{i,\min_D}}^{\tau_{i,g}} \mathbb{P}\{\xi^{chg}_{i,t'}\}}_{\textbf{D1) expected pulls from } t_g:\tau_{i,g} \textbf{ for changepoint detection}} \Bigg]$$

*Now, for **part A1** we obtain from from Theorem 1 in Auer et al. (2002a) that,*

$$\sum_{t'=N_{i,\min_E}}^{t_g} \mathbb{P}\{\xi_{i,t'}^{opt}\} \leq \sum_{t'=1}^{\infty}\sum_{n=1}^{t'} \mathbb{P}\left(\frac{\sum_{s=1}^{n} X_{i,s}}{n} \geq \mu_{i,g-1} + \frac{2\log(t)}{n}\right) + \sum_{t'=1}^{\infty}\sum_{q=1}^{t'} \mathbb{P}\left(\frac{\sum_{s=1}^{q} X_{i^*,s}}{q} \leq \mu_{i^*,g-1} - \frac{2\log(t)}{q}\right)$$
$$\overset{(a)}{\leq} \sum_{t'=1}^{\infty}\sum_{n=1}^{t'} \frac{1}{(t)^4} + \sum_{t'=1}^{\infty}\sum_{q=1}^{t'} \frac{1}{(t)^4} \overset{(b)}{\leq} \frac{\pi^2}{3}.$$

*where, (a) comes from the application of Chernoff-Hoeffding bound and (b) is obtained from Bazel's equation. Again, for **part C1** we can upper bound it as,*

$$N_{i,\min_D} + \max_{N_{1,t_g:\tau_g},\ldots,N_{K,t_g:\tau_g}} \left\{\sum_{i=1}^{K} N_{i,t_g:\tau_g} : N_{i,t_g:\tau_g} < N_{i,min_D}\right\}$$
$$\leq \frac{2\log(\frac{4t^2}{\delta})}{(\Delta_{i,g}^{chg})^2} + \max_{N_{1,t_g:\tau_g},\ldots,N_{K,t_g:\tau_g}} \left\{\sum_{i=1}^{K} N_{i,t_g:\tau_g} : N_{i,t_g:\tau_g} < N_{i,min_D}\right\}$$
$$\overset{(a)}{\leq} \frac{6\log(t)}{(\Delta_{i,g}^{chg})^2} + \max_{N_{1,t_g:\tau_g},\ldots,N_{K,t_g:\tau_g}} \left\{\sum_{i=1}^{K} N_{i,t_g:\tau_g} : N_{i,t_g:\tau_g} < N_{i,min_D}\right\}.$$

Here in (a) we substitute the value of $\delta = \frac{1}{t}$. Substituting these in $\mathbb{E}[N_{i,1:t}]$ we get,

$$\mathbb{E}[N_{i,1:t}] \overset{(a)}{\leq} \sum_{j=1}^{G} \Bigg[\underbrace{1 + \frac{8\log t}{(\Delta_{i,g}^{opt})^2} + \frac{\pi^2}{3}}_{\textbf{part A1}} + \underbrace{\frac{2\log(\frac{4t^2}{\delta})}{(\Delta_{i,g}^{chg})^2} + \sum_{t'=\tau_{i,g-1}+K+1}^{t_g} \delta}_{\textbf{part B1}} + \underbrace{\frac{2\log(\frac{4t^2}{\delta})}{(\Delta_{i,g}^{chg})^2} + \sum_{t'=\tau_{i,g}+K+1}^{\tau_{i,g}} \delta}_{\textbf{part D1}}$$
$$+ \underbrace{\frac{6\log(t)}{(\Delta_{i,g}^{chg})^2} + \max_{N_{1,t_g:\tau_g},\ldots,N_{K,t_g:\tau_g}} \left\{\sum_{i=1}^{K} N_{i,t_g:\tau_g} : N_{i,t_g:\tau_g} < N_{i,min_D}\right\}}_{\textbf{part C1}}\Bigg]$$
$$\overset{(b)}{\leq} \sum_{j=1}^{G} \Bigg[1 + \frac{8\log t}{(\Delta_{i,g}^{opt})^2} + \frac{\pi^2}{3} + \frac{6\log(t)}{(\Delta_{i,g}^{chg})^2} + \sum_{t'=\tau_{i,g-1}+K+1}^{t_g} \frac{1}{t}$$
$$+ \max_{N_{1,t_g:\tau_g},\ldots,N_{K,t_g:\tau_g}} \left\{\sum_{i=1}^{K} N_{i,t_g:\tau_g} : N_{i,t_g:\tau_g} < N_{i,min_D}\right\} + \frac{12\log(t)}{(\Delta_{i,g}^{chg})^2} + \sum_{t'=\tau_{i,g-1}+K+1}^{\tau_{i,g}} \frac{1}{t}\Bigg]$$

where, (a) comes from the result of Lemma 3 and in (b) we substitute $\delta = \frac{1}{t}$ from Remark 1.

Hence, the regret upper bound is given by summing over all arms in $\mathcal{A}$ and considering $\Delta_{i,g}^{opt} \leq \Delta_{\max,g+1}^{opt}$,

$$\mathbb{E}[R_t] \overset{(a)}{\leq} \sum_{i=1}^{K}\sum_{g=1}^{G} \left\{1 + \frac{8\log(t)}{\Delta_{i,g}^{opt}} + \frac{\pi^2}{3} + \frac{6\Delta_{i,g}^{opt}\log(t)}{(\Delta_{i,g}^{chg})^2} + 1 + \frac{2\Delta_{\max,g+1}^{opt}\log(\frac{4t^2}{\delta})}{(\Delta_{\epsilon_0,g})^2} + \frac{12\Delta_{i,g+1}^{opt}\log(t)}{(\Delta_{i,g}^{chg})^2} + 1\right\}$$
$$\overset{(b)}{\leq} \sum_{i=1}^{K}\sum_{g=1}^{G} \left\{3 + \frac{8\log(t)}{\Delta_{i,g}^{opt}} + \frac{\pi^2}{3} + \frac{6\Delta_{i,g}^{opt}\log(t)}{(\Delta_{i,g}^{chg})^2} + \frac{6\Delta_{\max,g+1}^{opt}\log(t)}{(\Delta_{\epsilon_0,g})^2} + \frac{12\Delta_{i,g+1}^{opt}\log(t)}{(\Delta_{i,g}^{chg})^2}\right\}.$$

Here, in (a) we use the Assumption 2, Assumption 3 and Discussion 1 to substitute $\Delta_g^{\epsilon_0}$ as the minimal gap that can be detected when $t_g > \tau_{i,g}$ and in (b) we substitute $\delta = \frac{1}{t}$

G Proof of Regret Bound of ImpCPD

Proof 5. *We proceed as like Theorem 3.1 in Auer and Ortner (2010) and we combine the changepoint detection sub-routine with this.*

Step 0.(Vital Assumption for proving): *In this proof, we assume that all the changepoints are detected by the algorithm with some delay. Since, all the gaps are significant (Assumption 3) and there is sufficient delay between two changepoints (Assumption 2), we can take this assumption for proving the result.*

Step 1.(Define some notations): *In this proof, we use $N_{i,g}$ instead of $N_{i,t_s:t_p}$ denoting the number of times the arm i is pulled between two restarts, ie, between t_s to t_p (as shown in algorithm) for the piece ρ_g. We define the confidence interval for the i-th arm as $S_{i,g} = \sqrt{\frac{\alpha \log(\psi T \epsilon_m^2)}{N_{i,g}}}$. The phase numbers are denoted by $m = 0, 1, \ldots, M$ where $M = \frac{1}{2}\log_{1+\gamma}\frac{T}{e}$. For each sub-optimal arm i, we consider their contribution to cumulative regret individually till each of the changepoint is detected. We also define $\mathcal{A}' = \left\{i \in \mathcal{A} : \Delta_{i,g}^{opt} \geq \sqrt{\frac{e}{T}}, \Delta_{i,g}^{chg} \geq \sqrt{\frac{e}{T}}, \forall g \in \mathcal{G}\right\}$.*

Step 2.(Define an optimality stopping phase $m_{i,g}$): *We define an optimality stopping phase $m_{i,g}$ for a sub-optimal arm $i \in \mathcal{A}'$ as the first phase after which the arm i is no longer pulled till the g-th changepoint is detected such that for $0 < \gamma \leq 1$,*

$$m_{i,g} = \min\left\{m : \sqrt{\alpha\epsilon_m} < \frac{\Delta_{i,g}^{opt}}{4(1+\gamma)^2}\right\}$$

Step 3.(Define a changepoint stopping phase $p_{i,g}$): *We define a changepoint stopping phase $p_{i,g}$ for an arm $i \in \mathcal{A}'$ such that it is the first phase when the changepoint g is detected such that for $0 < \gamma \leq 1$,*

$$p_{i,g} = \min\left\{m : \sqrt{\alpha\epsilon_m} < \frac{\Delta_{i,g}^{chg}}{4(1+\gamma)^2}\right\}$$

Step 4.(Regret for sub-optimal arm i being pulled after $m_{i,g}$-th phase): *Note, that on or after the $m_{i,g}$-th phase a sub-optimal arm $i \in \mathcal{A}'$ is not pulled anymore if these four conditions hold,*

$$\hat{\mu}_{i,g} < \mu_{i,g} + S_{i,g}, \qquad \hat{\mu}_{i*,g} > \mu_{i*,g} - S_{i*,g}, \qquad S_{i,g} > S_{i^*,g}, \qquad N_{i,g} \geq \ell_{m_{i,g}} \tag{5}$$

Also, in the $m_{i,g}$-th phase if $N_{i,g} \geq \ell_{m_{i,g}} = \frac{\log(\psi T \epsilon_{m_{i,g}}^2)}{2\epsilon_{m_{i,g}}}$ then we can show that,

$$S_{i,g} = \sqrt{\frac{\alpha \log(\psi T \epsilon_{m_{i,g}}^2)}{2N_{i,g}}} \leq \sqrt{\frac{\alpha \log(\psi T \epsilon_{m_{i,g}}^2)}{2\ell_{m_{i,g}}}} \leq \sqrt{\alpha\epsilon_{m_{i,g}} \frac{\log(\psi T \epsilon_{m_{i,g}}^2)}{\log(\psi T \epsilon_{m_{i,g}}^2)}} \leq \frac{\Delta_{i,g}^{opt}}{4(1+\gamma)^2}.$$

If indeed the four conditions in equation 5 hold then we can show that in the $m_{i,g}$-th phase,

$$\begin{aligned}
\hat{\mu}_{i,g} + S_{i,g} \leq \mu_{i,g} + 4(1+\gamma)^2 S_{i,g} - 2(1+\gamma)^2 S_{i,g} &\leq \mu_{i,g} + \Delta_{i,g}^{opt} - 2(1+\gamma)^2 S_{i,g} \\
&\leq \mu_{i*,g} - 2(1+\gamma)^2 S_{i^*,g} \\
&\leq \hat{\mu}_{i*,g} - S_{i^*,g}
\end{aligned}$$

Hence, the sub-optimal arm i is no longer pulled on or after the $m_{i,g}$-th phase. Therefore, to bound the number of pulls of the sub-optimal arm i, we need to bound the probability of the complementary of the four events in equation 6.

For the first event in equation 5, using Chernoff-Hoeffding bound we can upper bound the probability of the complementary of that event by,

$$\begin{aligned}
\sum_{m=0}^{m_{i,g}}\sum_{n=1}^{\ell_m}\mathbb{P}\left\{\frac{\sum_{s=1}^{n}X_{i,s}}{n} \geq \mu_{i,g}+\sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2n}}\right\} &\leq \sum_{m=0}^{m_{i,g}}\sum_{n=1}^{\ell_m}\exp\left(-2\left(\sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2n}}\right)^2 n\right)\\
&\leq \sum_{m=0}^{m_{i,g}}\sum_{n=1}^{\ell_m}\exp\left(-2\frac{\alpha\log(\psi T\epsilon_m^2)}{2n}n\right)\\
&\leq \sum_{m=0}^{m_{i,g}}\sum_{n=1}^{\ell_m}\frac{1}{\psi T\epsilon_m^2}\\
&\leq \sum_{m=0}^{m_{i,g}}\frac{\log(\psi T\epsilon_m^2)}{2\epsilon_m}\frac{1}{(\psi T\epsilon_m^2)^\alpha}\\
&\leq \frac{\log(\psi T\sum_{m=0}^{m_{i,g}}\epsilon_m^2)}{2(\psi T)^\alpha}\sum_{m=0}^{m_{i,g}}\frac{1}{\epsilon_m^{2\alpha+1}}\\
&\leq \frac{\log(\psi T)}{2(\psi T)^\alpha}\sum_{m=0}^{m_{i,g}}\frac{1}{\epsilon_m^{2\alpha+1}}.
\end{aligned}$$

Similarly, for the second event in equation 5, we can bound the probability of its complementary event by,

$$\begin{aligned}
\sum_{m=0}^{m_{i,g}}\sum_{n=1}^{\ell_m}\mathbb{P}\left\{\frac{\sum_{s=1}^{n}X_{i^*,s}}{n} \leq \mu_{i*,g}-\sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2n}}\right\} &\leq \sum_{m=0}^{m_{i,g}}\sum_{n=1}^{\ell_m}\exp\left(-2\left(\sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2n}}\right)^2 n\right)\\
&\leq \frac{\log(\psi T)}{2(\psi T)^\alpha}\sum_{m=0}^{m_{i,g}}\frac{1}{\epsilon_m^{2\alpha+1}}.
\end{aligned}$$

Also, for the third event in equation 5, we can bound the probability of its complementary event by,

$$\sum_{m=0}^{m_{i,g}}\mathbb{P}\{S_{i,g}<S_{i*,g}\}\leq\sum_{m=0}^{m_{i,g}}\mathbb{P}\{\hat{\mu}_{i,g}+S_{i,g}>\hat{\mu}_{i*,g}+S_{i*,g}\}$$

But the event $\hat{\mu}_{i,g}+S_{i,g}>\hat{\mu}_{i*,g}+S_{i*,g}$ is possible only when the following three events occur, $\{\hat{\mu}_{i*,g}\leq\mu_{i*,g}-S_{i*,g}\}\cup\{\hat{\mu}_{i,g}\geq\mu_{i,g}+S_{i,g}\}\cup\{\mu_{i*,g}-\mu_{i,g}<2(1+\gamma)^2S_{i,g}\}$. But the third event will not happen with high probability for $N_{i,g}\geq\ell_{m_{i,g}}$. Proceeding as before, we can show that the probability of the remaining two events is bounded by,

$$\begin{aligned}
\sum_{m=0}^{m_{i,g}} \mathbb{P}\{S_{i,g} < S_{i*,g}\} &\leq \sum_{m=0}^{m_{i,g}} \sum_{n=1}^{\ell_m} \sum_{q=1}^{\ell_m} \mathbb{P}\left\{ \frac{\sum_{s=1}^{n} X_{i,s}}{n} + \sqrt{\frac{\alpha \log(\psi T \epsilon_m^2)}{2n}} > \frac{\sum_{s=1}^{q} X_{i^*,s}}{q} + \sqrt{\frac{\alpha \log(\psi T \epsilon_m^2)}{2q}} \right\} \\
&\leq \sum_{m=0}^{m_{i,g}} \sum_{n=1}^{\ell_m} \mathbb{P}\left\{ \frac{\sum_{s=1}^{n} X_{i,s}}{n} \geq \mu_{i,g} + \sqrt{\frac{\alpha \log(\psi T \epsilon_m^2)}{2n}} \right\} \\
&+ \sum_{m=0}^{m_{i,g}} \sum_{q=1}^{\ell_m} \mathbb{P}\left\{ \frac{\sum_{s=1}^{q} X_{i^*,s}}{q} \leq \mu_{i^*,g} - \sqrt{\frac{\alpha \log(\psi T \epsilon_m^2)}{2q}} \right\} \\
&\leq \sum_{m=0}^{m_{i,g}} \sum_{n=1}^{\ell_m} \exp\left(-2 \frac{\alpha \log(\psi T \epsilon_m^2)}{2n} n \right) + \sum_{m=0}^{m_{i,g}} \sum_{q=1}^{\ell_m} \exp\left(-2 \frac{\alpha \log(\psi T \epsilon_m^2)}{2q} q \right) \\
&\leq \frac{\log(\psi T)}{2(\psi T)^{\alpha}} \sum_{m=0}^{m_{i,g}} \frac{1}{\epsilon_m^{2\alpha+1}}
\end{aligned}$$

Finally for the fourth event in equation 5, we can bound the probability of its complementary event by following the same steps as above,

$$\begin{aligned}
\sum_{m=0}^{m_{i,g}} \mathbb{P}\{N_{i,g} < \ell_{m_{i,g}}\} &\leq \sum_{m=0}^{m_{i,g}} \mathbb{P}\{\hat{\mu}_{i,g} + S_{i,g} < \hat{\mu}_{i*,g} + S_{i*,g}\} \\
&\leq \sum_{m=0}^{m_{i,g}} \sum_{n=1}^{\ell_m} \sum_{q=1}^{\ell_m} \mathbb{P}\left\{ \frac{\sum_{s=1}^{n} X_{i,s}}{n} + \sqrt{\frac{\alpha \log(\psi T \epsilon_m^2)}{2n}} < \frac{\sum_{s=1}^{q} X_{i^*,s}}{q} + \sqrt{\frac{\alpha \log(\psi T \epsilon_m^2)}{2q}} \right\} \\
&\leq \frac{\log(\psi T)}{(\psi T)^{\alpha}} \sum_{m=0}^{m_{i,g}} \frac{1}{\epsilon_m^{2\alpha+1}}.
\end{aligned}$$

Combining the above four cases we can bound the probability that a sub-optimal arm i will no longer be pulled on or after the $m_{i,g}$-th phase by,

$$\begin{aligned}
\frac{4\log(\psi T)}{(\psi T)^{\alpha}} \sum_{m=0}^{m_{i,g}} \frac{1}{\epsilon_{m_{i,g}}^{2\alpha+1}} &\leq \frac{4\log(\psi T)}{(\psi T)^{\alpha}} \sum_{m=0}^{M} \left(\frac{1}{\epsilon_m} \right)^{2\alpha+1} \\
&\overset{(a)}{\leq} \frac{4\log(\psi T)}{(\psi T)^{\alpha}} \left(\frac{(1+\gamma)((1+\gamma)^M - 1)}{(1+\gamma) - 1} \right)^{2\alpha+1} \\
&\overset{(b)}{\leq} \frac{4\log(\psi T)}{(\psi T)^{\alpha}} \left(\left(\frac{1+\gamma}{\gamma} \right) \sqrt{T} \right)^{2\alpha+1} \overset{(c)}{=} \frac{4\sqrt{T} C(\gamma, \alpha) \log(\psi T)}{(\psi)^{\alpha}}.
\end{aligned}$$

Here, in (a) we use the standard geometric progression formula, in (b) we substitute the value of $M = \frac{1}{2} \log_{1+\gamma} \frac{T}{e}$ and in (c) we substitute $C(\gamma, \alpha) = \left(\frac{1+\gamma}{\gamma} \right)^{2\alpha+1}$. Bounding this trivially by $T\Delta_{i,g}^{opt}$ for each arm $i \in \mathcal{A}'$ we get the regret suffered for all arm $i \in \mathcal{A}'$ after the $m_{i,g}$-th phase as,

$$\sum_{i \in \mathcal{A}'} \left(\frac{4T\Delta_{i,g}^{opt} \sqrt{T} C_{\gamma} \log(\psi T)}{(\psi)^{\alpha}} \right) = \sum_{i \in \mathcal{A}'} \left(\frac{4T^{\frac{3}{2}} C(\gamma, \alpha) \Delta_{i,g}^{opt} \log(\psi T)}{(\psi)^{\alpha}} \right) = \sum_{i \in \mathcal{A}'} \left(\frac{4C(\gamma, \alpha) \Delta_{i,g}^{opt} \log(\psi T)}{(\psi T^{-\frac{3}{2\alpha}})^{\alpha}} \right).$$

***Step 5.(Regret for pulling the sub-optimal arm** i **on or before** $m_{i,g}$**-th phase):** Either a sub-optimal arm gets pulled $\ell_{m_{i,g}}$ number of times till the $m_{i,g}$-th phase or after that the probability of it getting pulled is exponentially low (as shown in **step 4**). Hence, the number of times a sub-optimal arm i is pulled till the $m_{i,g}$-th phase is given by,*

$$N_{i,g} < \ell_{m_{i,g}} = \left\lceil \frac{\log(\psi T \epsilon_{m_{i,g}}^2)}{2\epsilon_{m_{i,g}}} \right\rceil$$

Hence, considering each arm $i \in \mathcal{A}'$ the total regret is bounded by,

$$\sum_{i \in \mathcal{A}'} \Delta_{i,g}^{opt} \left\lceil \frac{\log(\psi T \epsilon_{m_{i,g}}^2)}{2\epsilon_{m_{i,g}}} \right\rceil < \sum_{i \in \mathcal{A}'} \Delta_{i,g}^{opt} \left[1 + \frac{\log(\psi T \epsilon_{m_{i,g}}^2)}{2\epsilon_{m_{i,g}}} \right] \leq \sum_{i \in \mathcal{A}'} \Delta_{i,g}^{opt} \left[1 + \frac{8 \log(\psi T (\Delta_{i,g}^{opt})^4)}{(\Delta_{i,g}^{opt})^2} \right].$$

***Step 6.(Regret for not detecting changepoint** g **for arm** i **after the** $p_{i,g}$**-th phase):** First we recall a few additional notations, starting with $\hat{\mu}_{i,L_{p_{i,g-1}}:L_{m'}}$ denoting the sample mean of the i-th arm from $L_{p_{i,g-1}}$ to $L_{m'}$ timestep while $\hat{\mu}_{i,L_{m'}+1:L_{p_{i,g}}}$ denotes the sample mean of the i-th arm from $L_{m'}+1$ to $L_{p_{i,g}}$ timesteps where $m' = 1, \ldots, p_{i,g}$. Proceeding similarly as in **step 4** we can show that in the $p_{i,g}$-th phase for an arm $i \in \mathcal{A}'$ if $N_{i,L_{p_{i,g-1}}:L_m} \geq \ell_{p_{i,g}}$ then,*

$$S_i = \sqrt{\frac{\alpha \log(\psi T \epsilon_{p_{i,g}}^2)}{2N_{i,L_{p_{i,g-1}}:L_m}}} \leq \sqrt{\frac{\alpha \log(\psi T \epsilon_{p_{i,g}}^2)}{2\ell_{p_{i,g}}}} \leq \sqrt{\alpha \epsilon_{p_{i,g}} \frac{\log(\psi T \epsilon_{p_{i,g}}^2)}{\log(\psi T \epsilon_{p_{i,g}}^2)}} \leq \frac{\Delta_{i,g}^{chg}}{4(1+\gamma)^2}.$$

Furthermore, we can show that for a phase $m \in [1, p_{i,g}]$ if the following conditions hold for an arm $i \in \mathcal{A}'$ then the changepoint will definitely get detected, that is,

$$\begin{aligned}
\hat{\mu}_{i,L_{p_{i,g-1}}:L_m} + S_{i,N_{p_{i,g-1}}:L_m} &< \hat{\mu}_{i,L_m+1:L_{p_{i,g}}} - S_{i,L_m+1:L_{p_{i,g}}}, \\
\hat{\mu}_{i,L_{p_{i,g-1}}:L_m} - S_{i,L_{p_{i,g-1}}:L_m} &> \hat{\mu}_{i,L_m+1:L_{p_{i,g}}} + S_{i,L_m+1:L_{p_{i,g}}}, \\
S_{i,L_{p_{i,g-1}}:L_m} &< S_{i,L_m+1:L_{p_{i,g}}}, \\
N_{i,L_{p_{i,g-1}}:L_m} &\geq \ell_{p_{i,g}}.
\end{aligned} \tag{6}$$

Indeed, we can show that if there is a changepoint at g such that $\mu_{i,g-1} < \mu_{i,g}$ and if the first and third conditions in equation 6 hold in the $p_{i,g}$-th phase then,

$$\begin{aligned}
\hat{\mu}_{i,L_{p_{i,g-1}}:L_m} + S_{i,L_{p_{i,g-1}}:L_m} &\leq \mu_{i,g-1} + 4(1+\gamma)^2 S_{i,L_{p_{i,g-1}}:L_m} - 2(1+\gamma)^2 S_{i,L_{p_{i,g-1}}:L_m} \\
&\leq \mu_{i,g-1} + \Delta_{i,g}^{chg} - 2(1+\gamma)^2 S_{i,L_{p_{i,g-1}}:L_m} \\
&\leq \mu_{i,g} - 2(1+\gamma)^2 S_{i,L_m+1:L_{p_{i,g}}} \\
&\leq \hat{\mu}_{i,L_m+1:L_{p_{i,g}}} - S_{i,L_m+1:L_{p_{i,g}}}.
\end{aligned}$$

Conversely, for the case $\mu_{i,g-1} > \mu_{i,g}$ the second and third conditions will hold for the $p_{i,g}$-th phase. Now, to bound the regret we need to bound the probability of the complementary of these four events. For the complementary of the first event in equation 6 by using Chernoff-Hoeffding bound we can show that for all $m' \in [1, m]$,

$$\begin{aligned}
\sum_{m'=1}^{m} \mathbb{P}\{\hat{\mu}_{i,L_{p_{i,g-1}}:L_{m'}} \geq \mu_{i,g-1} + S_{i,L_{p_{i,g-1}}:L_{m'}}\} &\leq \sum_{m'=0}^{p_{i,g}} \sum_{n=1}^{\ell_{p_{i,g}}} \mathbb{P}\left\{\frac{\sum_{s=1}^{n} X_{i,s}}{n} \geq \mu_{i,g-1} + \sqrt{\frac{\alpha \log(\psi T \epsilon_{m'}^2)}{2n}}\right\} \\
&\leq \sum_{m'=0}^{p_{i,g}} \sum_{n=1}^{\ell_m} \exp\left(-2\frac{\alpha \log(\psi T \epsilon_{m'}^2)}{2n} n\right) \\
&\leq \sum_{m'=0}^{p_{i,g}} \sum_{n=1}^{\ell_m} \frac{1}{\psi T \epsilon_{m'}^2} \\
&\leq \sum_{m'=0}^{p_{i,g}} \frac{\log(\psi T \epsilon_{m'}^2)}{2\epsilon_{m'}} \frac{1}{(\psi T \epsilon_{m'}^2)^{\alpha}} \\
&\leq \frac{\log(\psi T)}{2(\psi T)^{\alpha}} \sum_{m'=0}^{p_{i,g}} \frac{1}{\epsilon_{m'}^{2\alpha+1}}.
\end{aligned}$$

Also, again by using Chernoff-Hoeffding bound we can show that for all $m' \in [m, p_{i,g}]$,

$$\begin{aligned}
\sum_{m'=m}^{p_{i,g}} \mathbb{P}\{\hat{\mu}_{i,L_{m'+1}:L_{p_{i,g}}} \leq \mu_{i,g} - S_{i,L_{m'+1}:L_{p_{i,g}}}\} &\leq \sum_{m'=0}^{p_{i,g}} \sum_{n=1}^{\ell_{p_{i,g}}} \mathbb{P}\left\{\frac{\sum_{s=1}^{n} X_{i,s}}{n} \geq \mu_{i,g} - \sqrt{\frac{\alpha \log(\psi T \epsilon_{m'}^2)}{2n}}\right\} \\
&\leq \sum_{m'=0}^{p_{i,g}} \sum_{n=1}^{\ell_m} \exp\left(-2\frac{\alpha \log(\psi T \epsilon_{m'}^2)}{2n} n\right) \\
&\leq \sum_{m'=0}^{p_{i,g}} \sum_{n=1}^{\ell_m} \frac{1}{\psi T \epsilon_{m'}^2} \\
&\leq \sum_{m'=0}^{p_{i,g}} \frac{\log(\psi T \epsilon_{m'}^2)}{2\epsilon_{m'}} \frac{1}{(\psi T \epsilon_{m'}^2)^{\alpha}} \\
&\leq \frac{\log(\psi T)}{2(\psi T)^{\alpha}} \sum_{m'=0}^{p_{i,g}} \frac{1}{\epsilon_{m'}^{2\alpha+1}}.
\end{aligned}$$

Hence, the probability that the changepoint is not detected by the first event in equation 6 is bounded by,

$$\frac{\log(\psi T)}{(\psi T)^{\alpha}} \sum_{m=0}^{p_{i,g}} \frac{1}{\epsilon_m^{2\alpha+1}}.$$

Similarly, we can bound the probability of the complementary of the second event in equation 6 for a $m' \in [1, m]$ or an $m' \in [m, p_{i,g}]$ as,

$$\begin{aligned}
&\sum_{m'=1}^{m} \mathbb{P}\{\hat{\mu}_{i,t_s:L_{m'}} \leq \mu_{i,g-1} - S_{i,t_s:L_{m'}}\} \leq \frac{\log(\psi T)}{2(\psi T)^{\alpha}} \sum_{m'=0}^{p_{i,g}} \frac{1}{\epsilon_{m'}^{2\alpha+1}}, \textbf{\textit{ and}} \\
&\sum_{m'=m}^{p_{i,g}} \mathbb{P}\{\hat{\mu}_{i,L_{m'+1}:p_{i,g}} \geq \mu_{i,g} + S_{i,L_{m'+1},p_{i,g}}\} \leq \frac{\log(\psi T)}{2(\psi T)^{\alpha}} \sum_{m'=0}^{p_{i,g}} \frac{1}{\epsilon_{m'}^{2\alpha+1}}.
\end{aligned}$$

Hence, the probability that the changepoint is not detected by the second event in equation 6 is again upper bounded by,

$$\frac{\log(\psi T)}{(\psi T)^{\alpha}} \sum_{m=0}^{p_{i,g}} \frac{1}{\epsilon_m^{2\alpha+1}}.$$

*Again, for the third event, following a similar procedure in **step 4**, we can bound its complementary event by,*

$$\sum_{m=0}^{p_{i,g}} \mathbb{P}\{S_{i,L_{p_{i,g-1}}:L_m} \geq S_{i,L_m+1:L_{p_{i,g}}}\} \leq 2\left(\frac{\log(\psi T)}{2(\psi T)^\alpha}\right) \sum_{m=0}^{p_{i,g}} \frac{1}{\epsilon_m^{2\alpha+1}}.$$

And finally, for the fourth event we can bound its complementary event by,

$$\sum_{m=0}^{p_{i,g}} \mathbb{P}\{N_{i,L_{p_{i,g-1}}:L_m} < \ell_{p_{i,g}}\} \leq 2\left(\frac{\log(\psi T)}{2(\psi T)^\alpha}\right) \sum_{m=0}^{p_{i,g}} \frac{1}{\epsilon_m^{2\alpha+1}}.$$

Combining the contribution from these four events, we can show that the probability of not detecting the changepoint g for the i-th arm after the $p_{i,g}$-th phase is upper bounded by,

$$\frac{2\log(\psi T)}{(\psi T \epsilon_{p_{i,g}}^2)^\alpha} \sum_{m=0}^{p_{i,g}} \frac{1}{\epsilon_{p_{i,g}}^{2\alpha+1}} \overset{(a)}{\leq} \frac{4C(\gamma,\alpha)\sqrt{T}\log(\psi T)}{(\psi)^\alpha}.$$

*Here, in (a) we substitute $C(\gamma,\alpha) = \left(\frac{1+\gamma}{\gamma}\right)^{2\alpha+1}$ and we follow the same steps as in **step 4** to reduce the expression to the above form. Furthermore, bounding the regret trivially (after the changepoint g) by $\Delta_{i,g+1}^{opt}$ for each arm $i \in \mathcal{A}'$, we get*

$$\sum_{i\in\mathcal{A}'}\left(\frac{2T\Delta_{i,g+1}^{opt}\sqrt{T}C(\gamma,\alpha)\log(\psi T)}{(\psi)^\alpha}\right) = \sum_{i\in\mathcal{A}'}\left(\frac{4T^{\frac{3}{2}}C(\gamma,\alpha)\Delta_{i,g+1}^{opt}\log(\psi T}{(\psi)^\alpha}\right) = \sum_{i\in\mathcal{A}'}\left(\frac{4C(\gamma,\alpha)\Delta_{i,g}^{opt}\log(\psi T)}{(\psi T^{-\frac{3}{2\alpha}})^\alpha}\right).$$

***Step 7.(Regret for not detecting a changepoint g for arm $i \in \mathcal{A}$ on or before the $p_{i,g}$-th phase):** The regret for not detecting the changepoint g on or before the $p_{i,g}$-th phase can be broken into two parts, **(a)** the worst case events from t_g to p_i and **(b)** the minimum number of pulls $\ell_{p_{i,g}}$ required to detect the changepoint. For the first part (a) we can use Assumption 2, Assumption 3, Discussion 1 and Definition 3 to upper bound the regret as,*

$$\sum_{i\in\mathcal{A}} \Delta_{\max,g+1}^{opt}\left\lceil\frac{8\log(\psi T\epsilon_{p_{i,g}}^2)}{(\Delta_{\epsilon_0,g})^2}\right\rceil < \sum_{i\in\mathcal{A}} \Delta_{\max,g+1}^{opt}\left[1+\frac{8\log(\psi T\epsilon_{p_{i,g}}^2)}{(\Delta_g^{\epsilon_0})^2}\right].$$

Again, for the second part (b) the arm $i \in \mathcal{A}'$ can be pulled no more than ℓ_{p_i} number of times. Hence, for each arm $i \in \mathcal{A}$ the regret for this case is bounded by,

$$\sum_{i\in\mathcal{A}'} \Delta_{i,g+1}^{opt}\left\lceil\frac{\log(\psi T\epsilon_{p_{i,g}}^2)}{2\epsilon_{p_{i,g}}}\right\rceil < \sum_{i\in\mathcal{A}'} \Delta_{i,g+1}^{opt}\left[1+\frac{\log(\psi T\epsilon_{p_{i,g}}^2)}{2\epsilon_{p_i,g}}\right].$$

Therefore, combining these two parts (a) and (b) we can show that the total regret for not detecting the changepoint till the $p_{i,g}$-th phase is given by,

$$\begin{aligned}
&\sum_{i\in\mathcal{A}}\left\{\Delta_{\max,g+1}^{opt} + \frac{8\Delta_{\max,g+1}^{opt}\log(\psi T\epsilon_{p_{i,g}}^2)}{(\Delta_{\epsilon_0,g})^2}\right\} + \sum_{i\in\mathcal{A}'}\left\{\Delta_{i,g+1}^{opt} + \frac{\Delta_{i,g+1}^{opt}\log(\psi T\epsilon_{p_{i,g}}^2)}{2\epsilon_{p_{i,g}}}\right\} \\
&\leq \sum_{i\in\mathcal{A}}\left\{\Delta_{\max,g+1}^{opt} + \frac{8\Delta_{\max,g+1}^{opt}\log(\psi T(\Delta_{i,g}^{chg})^4)}{(\Delta_{\epsilon_0,g})^2}\right\} + \sum_{i\in\mathcal{A}'}\left\{\Delta_{i,g+1}^{opt} + \frac{8\Delta_{i,g+1}^{opt}\log(\psi T(\Delta_{i,g}^{chg})^4)}{(\Delta_{i,g}^{chg})^2}\right\}
\end{aligned}$$

Step 8.(Final Regret bound): *Combining the assumption in* ***Step 0*** *and the problem definition, the expected regret till the T-th timestep is bounded by,*

$$\begin{aligned}\mathbb{E}[R_T] &= \sum_{i=1}^{K}\sum_{j=1}^{G}\Delta_{i,g}^{opt}\mathbb{E}[N_{i,g}] \\ &\leq \sum_{i\in\mathcal{A}'}\sum_{j=1}^{G}\Bigg[1+\underbrace{\frac{8C\left(\gamma,\alpha\right)\Delta_{i,g}^{opt}\log(\psi T)}{(\psi T^{-\frac{3}{2\alpha}})^{\alpha}}}_{\textit{from Step 4}}+\underbrace{\Delta_{i,g}^{opt}+\frac{8\log(\psi T(\Delta_{i,g}^{opt})^4)}{(\Delta_{i,g}^{opt})}}_{\textit{from Step 5}}+\underbrace{\frac{8C\left(\gamma,\alpha\right)\Delta_{i,g}^{opt}\log(\psi T)}{(\psi T^{-\frac{3}{2\alpha}})^{\alpha}}}_{\textit{from Step 6}}\Bigg] \\ &+\underbrace{\sum_{i\in\mathcal{A}}\sum_{j=1}^{G}\left[\Delta_{\max,g+1}^{opt}+\frac{8\Delta_{\max,g+1}^{opt}\log(\psi T(\Delta_{i,g}^{chg})^4)}{(\Delta_{\epsilon_0,g})^2}\right]+\sum_{i\in\mathcal{A}'}\sum_{j=1}^{G}\left[\Delta_{i,g+1}^{opt}+\frac{8\Delta_{i,g+1}^{opt}\log(\psi T(\Delta_{i,g}^{chg})^4)}{(\Delta_{i,g}^{chg})^2}\right]}_{\textit{from Step 7}}\end{aligned}$$

Now substituting the value of $\alpha=\frac{3}{2}$ and $\psi=\dfrac{T}{K^2\log K}$ in the above result, for the first part we get that,

$$\underbrace{\frac{8C\left(\gamma,\alpha\right)\Delta_{i,g}^{opt}\log(\psi T)}{(\psi T^{-\frac{3}{2\alpha}})^{\alpha}}}_{\textit{from Step 4}}=\frac{8C\left(\gamma,1.5\right)\Delta_{i,g}^{opt}\log(\frac{T^2}{K^2\log K})}{(\frac{T}{K^2\log K}T^{-1})^{\frac{3}{2}}}=\frac{16C_1\left(\gamma\right)\Delta_{i,g}^{opt}\log(\frac{T}{K\sqrt{\log K}})}{(K^2\log K)^{-\frac{3}{2}}}$$

where, $C_1(\gamma)=\left(\dfrac{1+\gamma}{\gamma}\right)^4$. For the second part we get,

$$\underbrace{\Delta_{i,g}^{opt}+\frac{8\log(\psi T(\Delta_{i,g}^{opt})^4)}{(\Delta_{i,g}^{opt})}}_{\textit{from Step 5}}=\Delta_{i,g}^{opt}+\frac{8\log(\frac{T}{K^2\log K}T(\Delta_{i,g}^{opt})^4)}{(\Delta_{i,g}^{opt})}=\Delta_{i,g}^{opt}+\frac{16\log(\frac{T(\Delta_{i,g}^{opt})}{K\sqrt{\log K}})}{(\Delta_{i,g}^{opt})}.$$

Similarly, for the third part, we get,

$$\underbrace{\frac{8C\left(\gamma,\alpha\right)\Delta_{i,g}^{opt}\log(\psi T)}{(\psi T^{-\frac{3}{2\alpha}})^{\alpha}}}_{\textit{from Step 6}}=\frac{16C_1\left(\gamma\right)\Delta_{i,g}^{opt}\log(\frac{T}{K\sqrt{\log K}})}{(K^2\log K)^{-\frac{3}{2}}}.$$

Finally, for the last part, we get that,

$$\begin{aligned}&\underbrace{\sum_{i\in\mathcal{A}}\sum_{j=1}^{G}\left[\Delta_{\max,g+1}^{opt}+\frac{8\Delta_{\max,g+1}^{opt}\log(\psi T(\Delta_{i,g}^{chg})^4)}{(\Delta_{\epsilon_0,g})^2}\right]+\sum_{i\in\mathcal{A}'}\sum_{j=1}^{G}\left[\Delta_{i,g+1}^{opt}+\frac{8\Delta_{i,g+1}^{opt}\log(\psi T(\Delta_{i,g}^{chg})^4)}{(\Delta_{i,g}^{chg})^2}\right]}_{\textit{from Step 7}} \\ &=\sum_{i\in\mathcal{A}}\sum_{j=1}^{G}\left[\Delta_{\max,g+1}^{opt}+\frac{16\Delta_{\max,g+1}^{opt}\log(\frac{T(\Delta_{i,g}^{chg})^2}{K\sqrt{\log K}})}{(\Delta_{\epsilon_0,g})^2}\right]+\sum_{i\in\mathcal{A}'}\sum_{j=1}^{G}\left[\Delta_{i,g+1}^{opt}+\frac{16\Delta_{i,g+1}^{opt}\log(\frac{T(\Delta_{i,g}^{chg})^2}{K\sqrt{\log K}})}{(\Delta_{i,g}^{chg})^2}\right].\end{aligned}$$

Hence, we get the main result of the regret bound.

H Proof of Corollary 1

Proof 6. *We first recall the result of Theorem 1 below,*

$$\mathbb{E}[R_t] \leq \sum_{i=1}^{K}\sum_{j=1}^{G}\Bigg\{\underbrace{3+\frac{8\log(t)}{\Delta_{i,g}^{opt}}+\frac{\pi^2}{3}}_{\textbf{part A}}+\underbrace{\frac{6\Delta_{i,g}^{opt}\log(t)}{(\Delta_{i,g}^{chg})^2}}_{\textbf{part B}}+\underbrace{\frac{6\Delta_{\max,g+1}^{opt}\log(t)}{(\Delta_{\epsilon_0,g})^2}}_{\textbf{part C}}+\underbrace{\frac{12\Delta_{i,g+1}^{opt}\log(t)}{(\Delta_{i,g}^{chg})^2}}_{\textbf{part D}}\Bigg\}.$$

Then, substituting $\Delta_{i,g}^{opt}=\Delta_{i,g}^{chg}=\Delta_g^{\epsilon_0}=\sqrt{\frac{K\log(T/G)}{\frac{T}{G}}}, \forall i\in\mathcal{A}, \forall g\in\mathcal{G}$ *we can show that for* ***part A****,*

$$KG\left\{\underbrace{3+\frac{8\log t}{(\Delta_{i,g}^{opt})}+\frac{\pi^2}{3}}_{\textbf{part A}}\right\}\leq 3KG+\frac{KG\pi^2}{3}+\frac{8K\sqrt{GT}\log T}{\sqrt{K\log T}}\leq 3KG+\frac{KG\pi^2}{3}+\frac{8\sqrt{KGT}\log T}{\sqrt{\log(T/G)}}$$
$$\leq 3KG+\frac{KG\pi^2}{3}+8\sqrt{KGT}\log T.$$

Similarly, for ***part B****, we can show that,*

$$KG\left\{\underbrace{\frac{6\Delta_{i,g}^{opt}\log(t)}{(\Delta_{i,g}^{chg})^2}}_{\textbf{part B}}\right\}\leq KG\frac{4\sqrt{\frac{GK\log(T/G)}{T}}\log(T)}{(\sqrt{\frac{GK\log(T/G)}{T}})^2}\leq 6\sqrt{GKT}\log T.$$

Again, for ***part C****, following the similar way as in* ***part B*** *we can show that,*

$$KG\left\{\underbrace{\frac{6\Delta_{\max,g+1}^{opt}\log(t)}{(\Delta_{\epsilon_0,g})^2}}_{\textbf{part C}}\right\}\leq 6\sqrt{GKT}\log T.$$

Finally, for ***part D*** *we can show that,*

$$KG\left\{\underbrace{\frac{12\Delta_{i,g+1}^{opt}\log(t)}{(\Delta_{i,g}^{chg})^2}}_{\textbf{part D}}\right\}\leq 12\sqrt{KGT}\log T.$$

Therefore, combining the four parts we get that the gap-independent regret upper bound of UCB-CPD is,

$$\mathbb{E}[R_T]\leq 3KG+\frac{KG\pi^2}{3}+32\sqrt{KGT}\log T.$$

I Proof of Corollary 2

Proof 7. *Again, we first recall the result of Theorem 2 below,*

$$\mathbb{E}[R_T] \leq \sum_{i\in\mathcal{A}'}\sum_{j=1}^{G}\Bigg[1+\underbrace{\frac{32C_1\left(\gamma\right)\Delta_{i,g}^{opt}\log(\frac{T}{K\sqrt{\log K}})}{(K^2\log K)^{-\frac{3}{2}}}}_{\textbf{\textit{part A}}}+\underbrace{\Delta_{i,g}^{opt}+\frac{16\log(\frac{T(\Delta_{i,g}^{opt})^2}{K\sqrt{\log K}})}{(\Delta_{i,g}^{opt})}}_{\textbf{\textit{part B}}}\Bigg]$$
$$+\underbrace{\sum_{i\in\mathcal{A}}\sum_{j=1}^{G}\left[\Delta_{\max,g+1}^{opt}+\frac{16\Delta_{\max,g+1}^{opt}\log(\frac{T(\Delta_{i,g}^{chg})^2}{K\sqrt{\log K}})}{(\Delta_{\epsilon_0,g})^2}\right]+\sum_{i\in\mathcal{A}'}\sum_{j=1}^{G}\left[\Delta_{i,g+1}^{opt}+\frac{16\Delta_{i,g+1}^{opt}\log(\frac{T(\Delta_{i,g}^{chg})^2}{K\sqrt{\log K}})}{(\Delta_{i,g}^{chg})^2}\right]}_{\textbf{\textit{part C}}}\Bigg]$$

Then, substituting $\Delta_{i,g}^{opt}=\Delta_{i,g}^{chg}=\Delta_{\epsilon_0,g}=\sqrt{\frac{K\log(T/G)}{\frac{T}{G}}},\forall i\in\mathcal{A},\forall g\in\mathcal{G}$ *and* $\gamma=0.05$ *such that* $C_1\left(\gamma\right)=C_1=9261$ *we can show that for* ***part A****,*

$$G\sum_{i\in\mathcal{A}'}\underbrace{\frac{32C_1\left(\gamma\right)\Delta_{i,g}^{opt}\log(\frac{T}{\sqrt{K^2\log K}})}{(K^2\log K)^{-\frac{3}{2}}}}_{\textbf{\textit{part A}}}\leq\frac{32C_1GK\sqrt{\frac{GK\log(T/G)}{T}}\log(\frac{T}{K\sqrt{\log K}})}{(K\log K)^{-1.5}}$$
$$\leq 32C_1G^{1.5}K^4(\log K)^{1.5}\sqrt{\log K}\frac{(\log T)^{1.5}}{\sqrt{T}}$$
$$\leq C_1G^{1.5}K^{4.5}(\log K)^2.$$

Again, for ***part B*** *we can show that,*

$$G\sum_{i\in\mathcal{A}'}\left(\underbrace{\Delta_{i,g}^{opt}+\frac{16\log(\frac{T(\Delta_{i,g}^{opt})^2}{K\sqrt{\log K}})}{(\Delta_{i,g}^{opt})^2}}_{\textbf{\textit{part B}}}\right)\leq GK\sqrt{\frac{GK\log(T/G)}{T}}+\frac{16GK\log(\frac{T}{K\sqrt{\log K}}(\sqrt{\frac{GK\log(T/G)}{T}})^2)}{(\sqrt{\frac{GK\log(T/G)}{T}})}$$
$$=G^{1.5}K\sqrt{\frac{K\log(T/G)}{T}}+\frac{16G\sqrt{KT}\log\left(\log TG\right)}{\sqrt{G\log T}}\overset{(a)}{\leq}G^{1.5}\sqrt{\frac{K^3\log(T/G)}{T}}+16\sqrt{GKT}.$$

where, (a) *comes from the identity that* $\frac{\log(\log(T))}{\sqrt{\log T}}\leq 1$ *and* $\frac{\log(\log G)}{\sqrt{\log T}}\leq 1$.

And finally, for ***part C****, similar to* ***part B****, we can show that,*

$$\underbrace{\sum_{i\in\mathcal{A}}\sum_{j=1}^{G}\left[\Delta_{\max,g+1}^{opt}+\frac{16\Delta_{\max,g+1}^{opt}\log(\frac{T(\Delta_{i,g}^{chg})^2}{K\sqrt{\log K}})}{(\Delta_{\epsilon_0,g})^2}\right]+\sum_{i\in\mathcal{A}'}\sum_{j=1}^{G}\left[\Delta_{i,g+1}^{opt}+\frac{16\Delta_{i,g+1}^{opt}\log(\frac{T(\Delta_{i,g}^{chg})^2}{K\sqrt{\log K}})}{(\Delta_{i,g}^{chg})^2}\right]}_{\textbf{\textit{part C}}}$$
$$\leq G^{1.5}\sqrt{\frac{K^3\log(T/G)}{T}}+16\sqrt{GKT}+G\sqrt{\frac{K^3\log(T/G)}{T}}+16\sqrt{GKT}$$
$$=2G^{1.5}\sqrt{\frac{K^3\log(T/G)}{T}}+32\sqrt{GKT}.$$

Hence, combining the four parts we get that that the gap-independent regret upper bound of ImpCPD is,

$$\mathbb{E}[R_T] \leq 3G^{1.5}\sqrt{\frac{K^3\log(T/G)}{T}} + C_1 G^{1.5}K^{4.5}(\log K)^2 + 48\sqrt{GKT}.$$

J Proof of Regret Lower Bound of Oracle Policy π^*

Proof 8. *We follow the same steps as in Theorem 2 of Maillard and Munos (2011) for proving the lower bound.*

Step 1.(Assumption): *The change of environment is not controlled by the learner and so the worst case scenario can be that the environment changes uniform randomly. A similar argument has also been made in the adaptive-bandit setting of Maillard and Munos (2011). So, let the horizon T be divided into G slices, each of length $\frac{T}{G}$. Hence, $T = G\frac{T}{G}$.*

Gap-dependent bound

Step 2.(Gap-dependent result): *From the stochastic bandit literature (Audibert and Bubeck, 2009), (Bubeck and Cesa-Bianchi, 2012), (Lattimore, 2015) we know that the gap-dependent regret bound for a horizon T and K arms in the stochastic bandit setting is lower bounded by,*

$$\mathbb{E}[R_T]_{SMAB-gap-dependent} \geq C\sum_{i=1}^{K}\frac{\log(\frac{T}{(H^{opt}})}{\Delta_i^{opt}}$$

where, C is a constant and $H^{opt} = \sum_{i=1}^{K}\frac{1}{\Delta_i^{opt}}$ is the optimality hardness which has been previously proposed in Audibert et al. (2010). Note, that the SMAB setting is a special case of the piecewise i.i.d setting where there is a single changepoint at $t_0 = 1$.

Step 3.(Regret for G changepoints): *The oracle policy π^* has access to the changepoints and is restarted without suffering any delay. Hence, combining Step 1 and Step 2 we can show that the gap-dependent regret lower bound scales as,*

$$\begin{aligned}\mathbb{E}[R_T]_{gap-dependent} &\geq \sum_{g=1}^{G}\left(C\sum_{i=1}^{K}\frac{\log(\frac{\frac{T}{G}}{(H^{opt,g}})}{\Delta_{i,g}^{opt}}\right)\\ &\geq C_1\sum_{g=1}^{G}\sum_{i=1}^{K}\frac{\log\frac{T}{GH_g^{opt}}}{\Delta_{i,g}^{opt}}\end{aligned}$$

where, C_1 is a constant and H_g^{opt} is the optimality hardness for the changepoint $g \in \mathcal{G}$ such that $H_g^{opt} = \sum_{i=1}^{K}\frac{1}{\Delta_{i,g}^{opt}}$.

Gap-independent bound

Step 4.(Reward of arms): *Let, for the interval ρ_g the optimal arm has a Bernoulli reward distribution of $\mu_{i^*,g} = Ber(\frac{(1+\epsilon_g)}{2})$ and all the other arms $i \in \mathcal{A}\setminus\{i_g^*\}$ have a Bernoulli reward distribution of $\mu_{i,g} = Ber(\frac{(1-\epsilon_g)}{2})$.*

Step 5.(Regret for G changepoints): *Let $I_g(t)$ denote the number of times a sub-optimal arm i is pulled between $t_g - 1$ timestep to t_g timestep. Now from Lemma 6.6 in Bubeck (2010) we know that for an ϵ_g of the order of $\sqrt{\frac{K}{s}}$ the following inequality holds,*

$$\sup_{i^*}\sum_{t=1}^{s}(\mu_{i^*,g} - \mu_{i_t,g}) \geq s\epsilon_g\left(1 - \frac{1}{K} - \sqrt{\frac{s\epsilon_g}{2K}\log\left(\frac{1-\epsilon_g}{1+\epsilon_g}\right)}\right).$$

In the adversarial setup, the adversary(or environment) chooses t_g such that T exactly divided into $\frac{T}{G}$ slices of G times, then the regret is lower bounded as,

$$\sup_{(i_g^*)_g} \sum_{g=1}^{G} \sum_{t=t_{g-1}}^{t_g} (\mu_{i^*,g} - \mu_{i_t,g}) \geq \sum_{g=1}^{G} \sum_{t=t_{g-1}}^{t_g} \epsilon_g \left(1 - \frac{1}{K} - \sqrt{\frac{I_g(t)\epsilon_g}{2K} \log\left(\frac{1-\epsilon_g}{1+\epsilon_g}\right)}\right)$$
$$\geq \sum_{g=1}^{G} \frac{T}{G} \epsilon_g \left(1 - \frac{1}{K} - \sqrt{\frac{T\epsilon_g}{2KG} \log\left(\frac{1-\epsilon_g}{1+\epsilon_g}\right)}\right).$$

The right hand side of the above expression can be optimized to yield that for any $\epsilon_g \approx \sqrt{\frac{T}{I_g(t)}}$,

$$\sup_{(i_g^*)_g} \sum_{g=1}^{G} \sum_{t=t_{g-1}}^{t_g} (\mu_{i^*,g} - \mu_{i_t,g}) \geq \frac{1}{20} \sum_{g=1}^{G} \sqrt{\frac{T}{KG}} = \frac{1}{20}\sqrt{KGT}.$$

Hence, the gap-independent regret bound for the oracle policy π^* is given by,

$$\mathbb{E}[R_T]_{gap-independent} \geq \frac{1}{20}\sqrt{KGT}.$$

K Proof of Regret Bound for OUCB1

Proposition 1. ***(Gap-independent Upper Bound for OUCB1)*** *The regret upper bound of OUCB1 for a horizon T and G changepoints is given by,*

$$\mathbb{E}[R_T] \leq C_1 \sqrt{GKT \log \frac{T}{G}}$$

where, C_1 is an integer constant.

Proof 9. ***Step 1.(Assumption):*** *The horizon T is divided into G changepoints such that T is divided into exactly $\frac{T}{G}$ slices where after every such slice the distribution changes for all $i \in \mathcal{A}$.*

Step 2.(Worst case result): *Now, we know from Auer et al. (2002a), Audibert and Bubeck (2009) that the worst case gap-independent regret bound of UCB1 is $C\sqrt{KT \log T}$ where C is an integer constant. But this is over the entire horizon T where there is no changepoint, or trivially one changepoint at the start at t_0.*

Step 3.(Oracle property): *So, combining Step 1 and Step 2, we can show that for any particular slice, the regret upper bound of OUCB1 is $C\sqrt{K(\frac{T}{G})\log(\frac{T}{G})}$. This is because it's an oracle UCB1, so it exactly restarts at the changepoints.*

Step 4.(Total Regret): *So, finally, the total regret bound for OUCB1 over entire horizon T is,*

$$\sum_{g=1}^{G} C\sqrt{K(\frac{T}{G})\log(\frac{T}{G})} = C_1\sqrt{GKT\log(\frac{T}{G})}.$$

where C_1 is an integer constant.

Discussion 9. *Hence, CUSUM and OUCB1 has the same gap-independent regret upper bound of $O(\sqrt{GT \log \frac{T}{G}})$.*

L Parameter Selection for Algorithms

For OUCB1 we use the same confidence interval as in Auer et al. (2002a) that is $\sqrt{\frac{2\log(t_p)}{N_{i,t_g:t_p}}}, \forall i \in \mathcal{A}$, where t_p is the current timestep and $N_{i,t_g:t_p}$ is the number of times an arm is pulled between two changepoints starting from $t_g, \forall g \in \mathcal{G}$

(as it is OUCB1). For DTS we use the discount factor $\gamma = 0.75$ as specified in Raj and Kalyani (2017) for slow varying environment. For DUCB we set $\gamma = 1 - \frac{1}{4}\sqrt{\frac{1}{T}}$ and for SWUCB we use the window size of $W = 4\sqrt{T \log T}$ as suggested in Garivier and Moulines (2011). Note, that for implementing the two passive algorithms we do not use the knowledge of the number of changepoints G to make these algorithms more robust, even though Garivier and Moulines (2011) suggests using $\gamma = 1 - \frac{1}{4}\sqrt{\frac{G}{T}}$ or $W = 4\sqrt{\frac{T \log T}{G}}$ for optimal performance. For CUSUM, we need to tune four parameters, M which is the minimum number of pulls required for each arm which must be first satisfied, parameter α which determines random uniform exploration between arms, and the CUSUM changepoint detection parameter h and tolerance factor ϵ. We choose $M = 40$, $\alpha = \sqrt{\frac{G}{T}\log\left(\frac{T}{G}\right)}$, $\epsilon = 0.1$ and $h = \log\left(\frac{T}{G}\right)$ (as suggested in Liu et al. (2017)) for both the experiments. Hence, CUSUM requires the knowledge of the number of changepoints in tuning its parameters. Similarly, Exp3.R requires tuning of three parameters δ, γ and H for changepoint detection. We choose $\delta = \sqrt{\frac{\log T}{KT}}$, $\gamma = \sqrt{\frac{K \log K \log T}{T}}$ and $H = \sqrt{T \log T}$ as suggested in Allesiardo et al. (2017). Finally, for UCB-CPD and ImpCPD, we choose the parameters as $\delta = \frac{1}{t}$ (Corollary 1) and $\gamma = 0.05$ (Corollary 2) respectively.