Abstract

We consider the setup of stochastic multi-armed bandits in the case when reward distributions are piecewise i.i.d. and bounded with unknown changepoints. We focus on the case when changes happen simultaneously on some (or 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 ($\Delta_{opt}^{g}$) and optimality-gaps ($\Delta_{opt}^{i,g}$). Under a slightly stronger set of assumptions, we show that as long as the compounded delayed detection for each changepoint is bounded there is no need for forced exploration to actively detect changepoints. We introduce two adaptations of UCB-strategies that employ scan-statistics in order to actively detect the changepoints, without knowing in advance the changepoints and also the mean before and after any change. Our first method UCBL-CPD does not know the number of changepoints $G$ or time horizon $T$ and achieves the first time-uniform concentration bound for this setting using the Laplace method of integration. The second strategy ImpCPD makes use of the knowledge of $T$ to achieve the order optimal regret bound of

$$\min \left\{ O\left( \sum_{i=1}^{K} \sum_{g=1}^{G} \frac{\log(T/H_{i,g})}{\Delta_{opt}^{i,g}} \right), O\left( \sqrt{GT} \right) \right\},$$

(where $H_{1,g}$ is the problem complexity) thereby closing an important gap with respect to the lower bound in a specific setting. Our theoretical findings are supported by numerical experiments on synthetic and real-life datasets.

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. The learning algorithm is provided with a set of decisions (or arms) which belong to the finite set $\mathcal{A}$ with individual arm indexed by $i$ such that $i = 1, \ldots, K$. 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 some (or all) arms changes at those changepoints. $G$ denotes the finite set of changepoints indexed by $g = 1, \ldots, G$, and $t_g$ denotes the time step when the changepoint $g$ occurs. At $t_g$ the learner detects $t_g$ with some delay. The reward distributions of all arms are unknown to the learner. 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. Moreover, in the health-care domain it is extremely risky to conduct additional uniform exploration of actions to detect simultaneous abrupt changes and relying only on the past history of interactions might lead to reliable detection of changepoints and safer policies.

A key aspect of this setting that the previous approaches have not fully utilized is that in the global changepoint setup when gaps are large enough on each arm so that they can be detected long before the next changepoint, no additional forced exploration is required to actively detect changepoints (like CUSUM (Liu et al., 2017), M-UCB (Cao et al., 2018)). But, this has to be done carefully as detecting a changepoint at $t_g$ requires some delay ($\tau_g$) and this might
endanger the detection of the next changepoint. This problem of **compounded delayed detection** becomes trickier if the successive changepoint gaps are not large enough and if they are not well separated (Figure 1). To avoid these complexities, past approaches have relied on just minimizing regret over a short window (passive algorithms) or conduct additional exploration regularly to detect changepoints and reset (active algorithms). These also inherently lead to less powerful and less informative worst case gap-independent bounds that depend on optimality gaps implicitly by assuming that gaps are evolving at a constant frequency, thus requiring the knowledge of \( G \) or \( T \) or both.

However, we can do better by utilizing three assumptions that the previous approaches take and reduce the regret incurred for forced explorations. These are assumption on global changepoint (1), a different separation of changepoint assumption (2) based on maximal delay, and a slightly stronger assumption (3) involving minimum detectable gaps (\( \Delta(t, \delta) \)). All the previous algorithms take some combination of these assumptions (see Table 1). But these only partially solve our problems, as specific care still needs to be taken by individual strategies as we have to show for each of our proposed detection strategies that their compounded delayed detection happens well before the next changepoint sets in. This approach provides the opportunity to study this interesting setting as an extension of SMAB. Hence, we can prove the gap-dependent regret upper bound that consist of both the changepoint gaps (\( \Delta_{\text{chg}}^{i,g} \)) and optimality gaps (\( \Delta_{\text{opt}}^{i,g} \)) for each changepoint \( g \in G \). Furthermore, now we will be able to give minimax (see 2.4.3 in Bubeck & Cesa-Bianchi (2012)) regret bounds that incorporates the Hardness factor \( H_{i,g} \) (introduced in Audibert et al. (2010), best-arm identification) which characterizes how difficult is the environment and depends on both \( \Delta_{\text{chg}}^{i,g} \) and \( \Delta_{\text{opt}}^{i,g} \). Also, for this setting we will be needing an additional hardness parameter \( H_{2,g} \) which captures the trade-off between optimality and changepoint gaps. We conjecture that an order optimal regret upper bound in the piecewise i.i.d setting of the order

\[
\min \left\{ O \left( \sum_{g=1}^{G} \sum_{i=1}^{K} \frac{H_{2,g} \log(T/(G H_{1,g}))}{\Delta_{\text{opt}}^{i,g}} \right), O \left( \sqrt{GT} \right) \right\}
\]

is attainable. Obtaining such optimal minimax bound for SMAB was discussed in Audibert & Bubeck (2009), Auer & Ortner (2010), Bubeck & Cesa-Bianchi (2012) and solved in Lattimore (2016). We further extend the results to piecewise i.i.d algorithms which is non-trivial given the changepoint and optimality gaps have to be tackled independently. Our contributions are mainly threefold: **Algorithmic**, **Time-Uniform bound**, and **Order-optimal Regret bound**.

1. **Algorithmic:** We propose two actively adaptive upper confidence bound (UCB) algorithms, referred to as UCB\(Laplace\)-Changepoint Detector (UCBL-CPD) and Improved Changepoint Detector (ImpCPD). Unlike CD-UCB, M-UCB and CUSUM, UCBL-CPD and ImpCPD do not conduct forced exploration to detect changepoints. They divide the time into slices, and for each time slice, for every arm they check the UCB, LCB mismatch based on past observations only to detect changepoints. While UCBL-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. Thus, controlling the compounded detection delay for ImpCPD is trickier than UCBL-CPD.

2. **Time-uniform bound:** The previous approaches CDUCB, CUSUM uses the Chernoff-Hoeffding inequality and M-UCB uses McDiarmids inequality with union bound to obtain the regret bound. DUCB and SWUCB uses the peeling argument which results in slightly tighter concentration bound but both these techniques results in less tight bounds than the Laplace method of integration used for UCBL-CPD.

   Moreover, UCBL-CPD has time-uniform bound as its confidence interval does not depend explicitly on \( t \) as opposed to other methods. On the contrary, ImpCPD which is not anytime and has access to \( T \) uses the usual union bound with geometrically increasing phase length (like peeling) is somewhat in between union and peeling argument. Both these proofs are of independent interest which can be used in other settings as well. A detailed comparison of union, peeling and Laplace bound can be found in Discussion 2.

3. **Regret bound:** We prove the gap-dependent regret upper bound that consist of both the changepoint gaps (\( \Delta_{\text{chg}}^{i,g} \)) and optimality gaps (\( \Delta_{\text{opt}}^{i,g} \)) for each changepoint \( g \in G \), \( i \in \mathcal{A} \) in Theorem 1, and 2 (under assumptions 1, 2, and 3). For the gap-independent result we show in the special case when all the gaps are same such that for all \( i \in \mathcal{A}, g \in G, \Delta_{\text{opt}}^{i,g} = \Delta_{\text{chg}}^{i,g} = \Delta(t, \delta), H_{1,g} = K(\Delta(t, \delta)^{-2} \text{ and } H_{2,g} = 1, \text{ UCBL-CPD and ImpCPD achieves } \sqrt{GT} \log T \text{ and } \sqrt{GT} \) respectively (Table 2, Corollary 1, and 2). Such a setup when all the gaps scale at least as \( \Omega(\sqrt{1/T}) \) is natural in view of existing works like Audibert & Bubeck (2009), Auer & Ortner (2010). We also prove a gap-dependent and independent lower bound for an optimal oracle policy \( \pi^* \) as opposed to Garivier & Moulines (2011) which proves a gap-independent lower bound for a policy \( \pi \) only when \( G = 2 \) (Theorem 3). Note, that an oracle policy \( \pi^* \) has access to the exact changepoints, where it is restarted with its past history of interactions erased. We show that UCBL-CPD and ImpCPD perform very well across diverse piecewise i.i.d environments (Section 6, and Appendix N).

The rest of the paper is organized as follows. We first setup

\[1\text{ requires a different separation of changepoint assumption.} \]
the problem 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, Section 7 contains related works, and Section 8 concludes the material is provided in this link [Link to supplementary].

2. Preliminaries

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 $\rho_g$ starts at $t_g$ and ends at $t_{g+1} - 1$. Let, $\mu_{i,g}$ be the expected mean of an arm $i$ for the piece $\rho_g$. 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 $\tau_g$ to $\tau_{g+1} - 1$ timesteps for any sequence of increasing $(\tau_g)_g$ of integers. Also, we define $\mu_{i, \tau_g, \tau_{g+1} - 1}$ as the empirical mean of the arm $i$ between $\tau_g$ to $\tau_{g+1} - 1$ timesteps. 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 $t \in \rho_g$.

**Definition 1.** 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 $\mu_{i,t-1} \neq \mu_{i,t}$, and $t^{*}_g \neq t^{*}_g$.

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

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

Thus, at each changepoint we assume that the optimal arm changes (other arms $i \in A \setminus \{i^{*}_{g-1}, i^{*}_g\}$ may or may not change), and the consecutive section $\rho_{g-1}$ and $\rho_g$ have two different optimal arms. Therefore, we make a distinction between the finite set of all arms $A$ and $A_{ch}^{g}$, such that $A_{ch}^{g}$ denotes only those arms $i \in A$ whose $\Delta_{i,g} > 0$ at the $t_g$-th changepoint.

Now, we carefully setup our problem. First, we define the minimum number of samples $n(t_g, \Delta, \delta)$ required for an arm $i \in A$ so that a deviation of $\Delta > 0$ of $\mu_{i,g-1}$ (or $\mu_{i,g}$) from $\mu_{i,g-1}$ (or $\mu_{i,g}$) before and after the $t_g$-th changepoint can be controlled with $(1 - \delta)$ probability. This is shown in Lemma 1. We then define the notion of maximal delay of a best policy $\pi^*$ in detecting a changepoint at $t_g$ starting exactly from $t_{g-1}$ (Lemma 2). Finally, we define the detectable changepoint gap $\Delta(t_g, \delta)$ based on the maximal delay of $\pi^*$ and $n(t_g, \Delta, \delta)$.

**Lemma 1.** (Control of large deviations) For our detection strategy using estimated means, it is sufficient to collect a minimum number of samples $n(t_g, \Delta, \delta) = \lceil \frac{1}{2} \log \left( \frac{2(1 - \delta)}{\delta} \right) \rceil$ for an arm $i \in A_{ch}^{g}$ before or after $t_g$ so that $|\mu_{i,g-1} - \mu_{i,g-1}| \leq \Delta$ or $|\mu_{i,g} - \mu_{i,g}| \leq \Delta$ with $(1 - \delta)$ probability and $t_{g-1} < t_g < t$.

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

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

\[
t_g + d_{\pi^*}(t_g - t_{g−1}) \leq t_g + \eta(t_g + 1 - t_g) = \eta t_{g+1} + (1 - \eta) t_g
\]

where \( \eta \in (0, 1) \), and \( d_{\pi^*}(t_g - t_{g−1}) \) is the maximal delay of a best detection strategy starting at \( t_{g−1} \).

Using Lemma 1 and Assumption 2 now we properly define detectable changepoint gap \( \Delta(t_g, \delta) \) to be such \( \Delta(t_g, \delta) \geq \sqrt{\log(2(x^2/\delta))/2x} \) where \( x = t_g + d_{\pi^*}(t_g - t_{g−1}) - t_{g−1} \).

**Lemma 2. (Detection Delay)** With the standard assumption that for each \( g \in G \), \( \Delta(t_g, \delta) \) scales atleast as \( \Omega(\sqrt{\log(1/t)}) \), then to detect \( \Delta(t_g, \delta) \) with probability \((1 - \delta)\), there exists a best detection strategy starting at \( t_{g−1} \) that suffers a maximum delay of,

\[
d_{\pi^*}(t_g - t_{g−1}) \leq \left( \frac{C(t, \delta, \eta)K \log(t_g^2)}{2(\Delta(t_g, \delta))^2} \right)
\]

where, \( C(t, \delta, \eta) \leq \eta \log(t/\delta) \), and \( \eta \in (0, 1) \).

**Definition 3.** The changepoint gap \( \Delta_{i,g}^{ch} \) for an arm \( i \in A_g \), \( g \in G \) is a \( \delta \)-optimal changepoint gap if \( \Delta_{i,g}^{ch} \geq \Delta(t_g, \delta) \).

**Assumption 3. (Minimum gap)** We assume that \( \forall g \in G \) the changepoint gaps \( \Delta_{i,g}^{ch}, \forall i \in A_g \) are \( \delta \)-optimal gaps.

**Discussion 1.** Thus, Assumption 2 makes sure that \( t_g + d_{\pi^*}(t_g - t_{g−1}) \) stays away from \( t_{g+1} \). This ensures that when restarting the detection strategy from \( t_g + d_{\pi^*}(t_g - t_{g−1}) \), the detection of \( t_{g+1} \) will not be too much endangered. Moreover, all the gaps \( \Delta_{i,g}^{ch}, \forall i \in A_g, \forall g \in G \) are \( \delta \)-optimal changepoint gaps following Assumption 3. A \( \delta \)-optimal changepoint gap requires a minimum sample of \( n(t_g, \Delta, \delta) \) to ensure that \( \Delta \) deviation from the mean occur with \((1 - \delta)\) probability (Lemma 1). A reasonable detection strategy \( \pi \) has its maximal delay \( d_{\pi}(t_g - t_{g−1}) \geq d_{\pi^*}(t_g - t_{g−1}) \) as it only has observation to detect \( t_g \) from \( t_{g−1} \) and not from \( t_{g−1} \) (unlike \( \pi^* \)). It tries to minimize \( d_{\pi}(t_g - t_{g−1}) \) in detecting \( t_g \) such that \( d_{\pi}(t_g - t_{g−1}) \leq (1 + \beta)d_{\pi}(t_g - t_{g−1}) \) holds with high probability for some \( \beta \in (0, 1) \) and \( \eta \in (0, 1) \). This approach is different than 

**Definition 4.** Let the optimality gap \( \Delta_{i,g}^{opt} \) for an arm \( i_{t'} \neq i_{t''} \), \( \forall t' \in [t_{g−1}, t_g - 1] \) be defined as, \( \Delta_{i,g}^{opt} = \mu_{\pi^*} - \mu_{\pi} \).

**3. Problem Formulation**

**Regret Definition:** The objective of the learner is to minimize the cumulative regret till \( T \), which is defined as:

\[
R_T = \sum_{t' = 1}^{T} \mu_{\pi_t^*} - \sum_{t' = 1}^{T} \mu_{\pi_t} \mathbb{I}\{i_{t'} \neq i_{t'}^*\}
\]

where \( T \) is the horizon, \( \mu_{\pi_t^*} \) is the expected mean of the optimal arm at the \( t' \) timestep and \( \mu_{\pi_t} \mathbb{I}\{i_{t'} \neq i_{t'}^*\} \) is the expected mean of the arm chosen by the learner at the \( t' \) timestep when it was not the optimal arm \( i_{t'}^* \). The expected regret of an algorithm after \( T \) rounds can be written as,

\[
\mathbb{E}[R_T] = \mathbb{E}\left[ \sum_{t' = 1}^{T} \mu_{\pi_t^*} - \sum_{t' = 1}^{T} \mu_{\pi_t} \mathbb{I}\{i_{t'} \neq i_{t'}^*\} \right]
\]

where \( \mathbb{E} \) is from Assumption 1, and \( \mathbb{E} \) from Definition 4.

**Problem Complexity:** We define the hardness of a changepoint \( g \in G \) using optimality and changepoint gaps by modifying the definitions of **problem complexity** as introduced in Audibert et al. (2010) for stochastic bandits:

\[
H_{1,g} = \max_{i,g} \frac{1}{\sum_{i,g} \Delta_{i,g}^{opt} \left( \sum_{i,g} \Delta_{i,g}^{opt} \right)^{1/2}} \leq \frac{K}{(\Delta_{i,g}^{opt})^2} \leq \frac{K}{(\Delta_{i,g}^{opt})^2} H_{2,g} = \frac{\Delta_{1,g}}{\Delta_{max,g+1}} H_{2,g}
\]

where, \( \Delta_{max,g+1} = \max_{i \in A_g} \Delta_{i,g}^{opt} \). The hardness parameter \( H_{2,g} \) captures the tradeoff between the minimum detectable gap \( \Delta(t_g, \delta) \) and maximum optimality gap of the next changepoint \( \Delta_{max,g+1} \) which serves as an upper bound to all such possible trade-offs at changepoint \( g \). The relation between the above complexity terms can be derived as:

\[
H_{2,g} \leq H_{1,g} - \frac{K}{(\Delta_{i,g}^{opt})^2} H_{2,g}
\]

Note that, \( \frac{\Delta_{i,g}^{opt}}{\Delta_{i,g}^{ch}} \leq H_{2,g} \), \( \forall i \in A_g \), \( g \in G \). In the special case when \( \Delta_{i,g}^{opt} = \Delta_{i,g}^{ch} \), \( \forall i \in A_g, g \in G \), then,

\[
H_{1,g} = K(\Delta(t_g, \delta))^{-2}, \quad H_{2,g} = 1
\]

**4. Algorithms**

We first introduce the policy **UCBL-CPD** in Algorithm 1 which is an adaptive algorithm based on the standard **UCB1** (Auer et al., 2002a) approach. **UCBL-CPD** pulls an arm at
every timestep as like UCB1 but has the time-uniform concentration bound that holds simultaneously for all timestep $t$. It calls upon the Changepoint Detection (CPD) subroutine in Algorithm 2 for detecting a changepoint. Note, that unlike CD-UCB, CUSUM, and M-UCB the UCBL-CPD does not conduct forced exploration to detect changepoints. UCBL-CPD is an anytime algorithm which does not require the horizon as an input parameter or to tune its parameter $\delta$. This is in stark contrast with CD-UCB, CUSUM, M-UCB, DUCB or SWUCB, that require the knowledge of $G$ or $T$ for optimal performance. we define the confidence interval of UCB-Laplace as follows:

$$S_{t,s,t_p} := \sqrt{\left(1 + \frac{1}{N_{t,s,t_p}}\right) \log\left(\frac{\sqrt{N_{t,s,t_p}} + 1/\delta}{2N_{t,s,t_p}}\right)}$$

**Algorithm 1** UCB Laplace (UCBL-CPD)

1. **Input:** $\delta > 0$; $S_{t,s,t_p}$ from (1)
2. **Initialization:** $t_s := 1$, $t_p := 1$.
3. Pull each arm once.
4. for $t = K + 1, \ldots, T$ do
5. Pull arm $j \in \arg \max_{i \in A} \{\hat{\mu}_{i,t,s,t_p} + S_{t,s,t_p}\}$, observe reward $X_{j,t}$.
6. Update $\hat{\mu}_{j,t,s,t_p}$, $N_{j,t,s,t_p} := N_{j,t,s,t_p} + 1$.
7. $t_p := t_p + 1$.
8. if (CPD($t_s$, $t_p$, $\delta$)) then
9. **Restart:** Set $\hat{\mu}_{i,t,s,t_p} := 0$, $N_{i,t,s,t_p} := 0$, $\forall i \in A$, $t_s := t$, $t_p := t_s$.
10. Pull each arm once.
11. **end if**
12. **end for**

**Algorithm 2** Changepoint Detection($t_s$, $t_p$, $\delta$) (CPD)

1. for $i = 1, \ldots, K$ do
2. for $t' = t_s, \ldots, t_p$ do
3. if $(\hat{\mu}_{i,t,s,t'} + S_{i,t,s,t'} < \hat{\mu}_{i,t'+1,t_s} - S_{t,t'+1,t_p})$ or $(\hat{\mu}_{i,t,s,t'} - S_{i,t,s,t'} > \hat{\mu}_{i,t'+1,t_s} + S_{t,t'+1,t_p})$ then
4. Return True.
5. **end if**
6. **end for**
7. **end for**

We introduce the phase-based ImpCPD in Algorithm 3 in Appendix D. ImpCPD calls upon the changepoint detector CPDI only at the end of phases and saves upon computation time without incurring additional regret.

**Running time of algorithms:** UCBL-CPD (like CD-UCB, and CUSUM) calls the changepoint detection at every timestep, and ImpCPD calls upon the sub-routine only at end of phases. Hence, for a fixed horizon $T$, $K$ arms, UCBL-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 designing ImpCPD carefully and appropriately modifying the confidence interval, this reduction comes at no additional cost in the order of regret (see Discussion 7)

**5. Main Results**

**Lemma 3.** (Control of bad-event by Laplace method) Let, $N_{i,t,s,t}$ be the number of times an arm $i$ is pulled from $t_s$ till the $t$-th timestep such that $t > t_p$, then at the $t$-th timestep for all $\delta \in (0, \frac{1}{2})$ it holds that,

$$\mathbb{P}\left\{\xi_{t,i,s}^{chg} \geq \delta\right\} \leq 2\delta$$

where the event $\xi_{t,i,s}^{chg} = \{\hat{\mu}_{i,t,s,t'} - S_{i,t,s,t'} > \hat{\mu}_{i,t'+1,t} + S_{t,t'+1,t_p}\}$. Then the event $\xi_{t,i,s}^{chg}$ using $M^\Lambda$. The proof is in Appendix E.

**Remark 1.** Choosing $\delta = \frac{1}{T}$, in Lemma 3 we can show that $\mathbb{P}\left\{\xi_{t,i,s}^{chg} \geq \frac{\delta}{T}\right\} \leq \frac{2}{T}$, where the event $\xi_{t,i,s}^{chg}$ is the bad event. Note, that $\delta$ does not depend on the knowledge of horizon $T$.

**Discussion 2.** Time-uniform bounds depends only on the number of pulls $N_{i,t,s,t'}$ and implicitly on $t$ based on the parameter $\delta$. The concentration bounds based on peeling and union bound method depends explicitly on $t$ and has larger coefficients attached to them. In Table 3 we give a comparison over the three concentration bound method involving Union bound, Peeling and Laplace method. We provide the proof of the construction of concentration bound for our changepoint detection strategy by union bound in Lemma 4 in Appendix F for completeness. The log log$(t)$ scaling of the peeling method is not better than the one derived by the Laplace method, unless for huge timestep $t$ ($t > 10^6$, for $\delta = 0.05$ and any $\alpha > 1$). Laplace method uses the sub-Gaussian nature of the variables to give such sharp concentration bounds as opposed to other methods.

**Theorem 1.** (Gap-dependent bound of UCBL-CPD) For $\eta \geq \frac{6}{\log(t'+1)}$, $\delta = \frac{1}{T}$, the expected cumulative regret of


**Proof 3. (Outline)** For each \( g \in \mathcal{G} \) we bound the probability of the bad event of discarding the optimal arm between changepoints \( g - 1 \) and \( g \) and not detecting the changepoint using the Laplace method (see Lemma 3). The proof of Theorem 1 is given in Appendix G. Note, that for a changepoint \( g \), compounded detection delay \( d_\pi(t_g - t_{g-1}) \geq d_\pi(t_g - t_{g-1}) \) as it lacks observations from \( t_g \) itself. But it suffices to show that at least \( d_\pi(t_g - t_{g-1}) = d_\pi(t_g - t_{g-1} + \delta \) will not be much greater than \( d_\pi(t_g - t_{g-1}) \) and for each \( g \in \mathcal{G} \) it is bounded with high probability as long as the \( \eta \) separation is maintained by Assumption 2 (Step 6).

**Discussion 3.** In Theorem 1, (a) is the regret suffered for finding the optimal arm between changepoints \( g - 1 \) and \( g \), (b) is the maximal regret for delayed detection of \( g \), and (c) is the regret suffered for total compounded delayed detection.

**Theorem 2. (Gap-dependent bound of ImpCPD )** For \( \eta \geq \frac{8}{2^{2K} + 1} \), the expected cumulative regret of ImpCPD using CPD is upper bounded by:

\[
\mathbb{E}[R_T] \leq \sum_{g=1}^{G} \sum_{i \in A} \left[ 48KC (\gamma) \frac{\Delta_{i,g}^{\text{opt}}}{\log\left(\frac{T}{K \sqrt{\log K}}\right)} + \frac{16 \log \left(\frac{\Delta_{i,g}^{\text{opt}}}{\log\left(\frac{T}{K \sqrt{\log K}}\right)}\right)}{(\Delta_{i,g}^{\text{cha}})^2} + 16H_{2,g} \log \left(\frac{T(\Delta_{i,g}^{\text{cha}})^2}{K \log K}\right) \right]
\]

**Proof 4. (Outline)** The key to proving this theorem is to carefully construct each geometrically increasing phase length \( \ell_m \) so that the probability not pulling the optimal arm between two changepoints \( g - 1 \) to \( g \) is bounded. Simultaneously, we use the phase length \( \ell_m \), confidence interval \( S_{i,t;\gamma} \), and exploration factor \( \gamma \) to control the bad event of not detecting the changepoint \( g \). We use Chernoff-Hoeffding inequality to bound the probability of the bad events. We have to further balance \( \ell_m \), and \( S_{i,t;\gamma} \) by carefully defining \( \psi \) so that \( \gamma \) is small enough and CPDI is called more often. We also have to use additional union bounds to control the event that arms are getting pulled unequal number of times within each phase length. The proof is in Appendix H.

**Discussion 4.** In Theorem 2, (a) is the regret suffered for calling the CPDI only at end of phases, (b) is the regret for finding the optimal arm between changepoints \( g - 1 \) and \( g \), (c) is the regret for delayed detection of \( g \), and (d) is the regret suffered for total compounded delayed detection.

**Discussion 5.** From Theorem 1 and 2 we see that UCBL-CPD and ImpCPD perform better than a reasonable detection strategy (see Lemma 2) as our proposed algorithms have the coefficient in their compounded detection delay of order \( O(K) \) that is less than order \( O(\sqrt{T} \log T) \) of \( C(t, \delta, \eta) \). This is reasonable as in the bandit setup each arm \( i \in A \) might be pulled a logarithmic number of times before detecting a changepoint.

**Corollary 1. (Gap-independent bound of UCBL-CPD )** In the specific scenario, when all the gaps are same, that is \( \Delta_{i,g}^{\text{opt}} = \Delta_{i,g}^{\text{cha}} = \Delta = \sqrt{\frac{K \log(\frac{T}{G})}{T/G}} \), then the worst case gap-independent regret bound of UCBL-CPD is given by:

\[
\mathbb{E}[R_T] \leq O(\sqrt{GT} \log T).
\]

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

**Discussion 6.** In Corollary 1, the largest contributing factor to the gap-independent regret of UCBL-CPD is of the order \( O(\sqrt{GT} \log T) \), same as that of DUCB but weaker than CUSUM and SWUCB. The additional \( O(\log T) \) factor is the cost UCBL-CPD must pay for not knowing \( T \) and \( G \).

**Corollary 2. (Gap-independent bound of ImpCPD )** In the specific scenario, when all the gaps are same, that is \( \Delta_{i,g}^{\text{opt}} = \Delta_{i,g}^{\text{cha}} = \Delta = \sqrt{\frac{K \log(\frac{T}{G})}{T/G}} \), then the worst case gap-independent regret bound of ImpCPD is given by:

\[
\mathbb{E}[R_T] \leq C_1 G^{1.5} K^{1.5} (\log K)^2 + O(\sqrt{GT})
\]

where \( C_1 \) is an integer constant.

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

**Discussion 7.** In Corollary 2, the largest contributing factor to the gap-independent regret of ImpCPD is of the order...
We compare UCBL-CPD and ImpCPD against Oracle when all the gaps are same such that for all \( \pi \)

\[
\Delta \in \Omega \left( \sqrt{GT} \right).
\]

This is lower than the regret upper bound of DUCB, SWUCB, EXP3.R and CUSUM (Table 1). The smaller the value of the exploration parameter \( \gamma \) the larger is the constant \( C_1 \) associated with the factor \( GK^{4.5}(\log K)^2 \).

Now, \( \gamma \) determines how frequently CPDI is called by ImpCPD and by modifying the confidence interval and phase-length we have been able to control the probability of not detecting the changepoint at the cost of additional regret that only scales with \( K \) and not with \( T \).

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

\[
\mathbb{E}_{\pi^*}[R_T] \geq \min \left\{ \Omega \left( \frac{\sum_{g=1}^{K} \sum_{i=1}^{K} \log \left( \frac{T}{(GH_{1,g})} \right)}{\Delta_{i,g}^{opt}} \right), \Omega \left( \sqrt{GT} \right) \right\}
\]

where, \( H_{1,g} = \sum_{i=1}^{K} \frac{1}{(\Delta_{i,g}^{opt})^2} \) is the hardness of the problem.

**Proof 7.** The key observation to prove this theorem is that the worst case scenario can occur when environment changes uniform randomly. A similar argument has also been made in the adaptive-bandit setting of Maillard & Munos (2011). So, let the horizon \( T \) be divided into \( G \) slices, each of length \( (T/G) \). For each of these slices an oracle algorithm using OCUCB (Lattimore, 2015) should get the optimal SMAB regret without suffering any delay. The proof is in Appendix K.

**Discussion 8.** This lower bound is weaker than the bound proposed in Wei et al. (2016) as they do not require the knowledge of \( G \) or \( T \). But we provide this for completion to discuss oracle-based bounds and also because the previous approaches do not touch upon this approach. For a non-oracle policy the additional trade-off between the changepoint gap and the next optimality gap is captured by \( H_{2,g} \). As long as the delayed detection is bounded with high probability we should get a similar scaling for a good detection algorithm minimizing regret for each of these slices of length \( (T/G) \). ImpCPD which has a gap-independent regret upper bound of \( O(\sqrt{GT}) \) reaches the lower bound of the policy \( \pi^* \) in an order optimal sense. Also, in the special case when all the gaps are same such that for all \( i \in \mathcal{A}, g \in \mathcal{G} \), \( \Delta_{i,g}^{opt} = \Delta(t_g, \delta) = \Delta_{i,g}^0 \), \( H_{1,g} = K(\Delta(t_g, \delta))^{-2} \) and \( H_{2,g} = 1 \), then ImpCPD with a gap-dependent bound of \( O\left( \sum_{g=1}^{G} \sum_{i=1}^{K} \frac{\log(T/H_{1,g})}{\Delta_{i,g}^{opt}} \right) \) matches the gap dependent lower bound of \( \pi^* \) except the factor \( G \) in the log term (ignoring the \( \log(\frac{1}{\log K}) \) term).

### 6. Experiments

We compare UCBL-CPD and ImpCPD against Oracle Thompson Sampling (OTS), EXP3.R, Discounted Thompson Sampling (DTS), Discounted UCB (DUCB), Sliding Window UCB (SWUCB), Monitored-UCB (M-UCB) and CUSUM-UCB (CUSUM) in four 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 M and additional experiments are shown in Appendix N.

**Experiment 1 (Bernoulli 3 arms):** This experiment is conducted to test the performance of algorithms in Bernoulli distribution over a short horizon \( T = 4000 \) and and small number of arms \( K = 3 \). There are 3 changepoints in this testbed and the expected mean of the arms changes as shown in equation 2. The experiment is shown in Figure 2(a) where we can clearly see that UCBL-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. UCBL-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 UCBL-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.

\[
\begin{align*}
\begin{array}{c}
r_1 = 0.1, r_2 = 0.2, r_3 = 0.9 \\
r_1 = 0.4, r_2 = 0.9, r_3 = 0.1 \\
r_1 = 0.5, r_2 = 0.1, r_3 = 0.2 \\
r_1 = 0.2, r_2 = 0.2, r_3 = 0.3
\end{array}
\end{align*}
\]

**Experiment 2 (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 use Jester 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. 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 uniformly randomly a single algorithm may learn quicker across users if all the users prefer a few common items. Note, that we violate Assumption 2 and Assumption 3 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. ImpCPD and UCBL-CPD is only able to detect 2 of the 6 changepoints and restart while CUSUM failed to detect any of the changepoints. Note that UCBL-CPD and ImpCPD performs slightly worse than DUCB 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.

![Image](image_url)

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

(b) Expt-2: 20 Users, 100 items, Rank 3 approximation of Jester Dataset

Figure 2. Cumulative regret of algorithms in Jester dataset

7. Related Works

Previous algorithms can be broadly divided into passive and actively adaptive algorithms. Passive algorithms like Discounted UCB (DUCB) (Kocsis & Szepesvári, 2006), Sliding Window UCB (SWUCB) (Garivier & Moulines, 2011) and Discounted Thompson Sampling (DTS) (Raj & Kalyani, 2017) do not actively try to detect changepoints and thus perform badly when changepoints are of large magnitude and are well-separated. 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 algorithms CD-UCB (Liu et al., 2017), CUSUM (Liu et al., 2017) and M-UCB (Cao et al., 2018) rely on additional forced exploration for changepoint detection. With probability they employ some changepoint detection mechanism or pull the arm with highest UCB with probability (exploitation). This is hard to tune in experiments and come with limited theoretical guarantees. CD-UCB requires that the exploration parameter is set to for proving theoretical guarantees. CUSUM (Liu et al., 2017) performs a two-sided CUSUM test to detect changepoints and it empirically outperforms CD-UCB. CUSUM requires the knowledge of and for tuning and its theoretical guarantees only hold for Bernoulli rewards for widely separated changepoints. Also, CUSUM wrongly applies Hoeffding inequality to a random number of pulls (see eq (31), (32) in Liu et al. (2017)) which raises serious concerns about the validity of the rest of their analysis. Finally, M-UCB also requires the knowledge of and for theoretical guarantees. Another approach involves the Generalized Likelihood Ratio Test which was recently studied by Maillard (2019) and Besson & Kaufmann (2019). This is a different approach than others and looks at the ratio of the likelihood of the sequence of rewards coming from two different distributions and calculates the sufficient statistics to detect changepoints. An extended discussion can be found in Appendix A.

8. Conclusions and Future Works

We studied the piecewise i.i.d environment under Assumption 1, 2 and 3 such that actively adaptive algorithms do not need to conduct forced exploration to detect changepoints. We studied two UCB algorithms, UCBL-CPD and ImpCPD which are adaptive and restarts once the changepoints are detected. We derived the first gap-dependent bounds for these actively adaptive algorithms incorporating the hardness factor. The anytime UCBL-CPD uses the Laplace method to derive sharp concentration bound, and ImpCPD achieves the order optimal regret bound which is an improvement over all the existing algorithms (in a specific setting). Empirically, they perform very well in various environments and is only outperformed by oracle algorithms. Future works include incorporating the knowledge of localization in these adaptive algorithms.
References


