Safety Aware Changepoint Detection for Piecewise i.i.d. Bandits

Subhojyoti Mukherjee[1]

[1]Electrical & Computer Engineering Department, UW-Madison, Madison, Wisconsin, USA

Abstract

In this paper, we consider the setting of piecewise i.i.d. bandits under a safety constraint. In this piecewise i.i.d. setting, there exists a finite number of changepoints where the mean of some or all arms change simultaneously. We introduce the safety constraint studied in Wu et al. [2016] to this setting such that at any round the cumulative reward is above a constant factor of the default action reward. We propose two actively adaptive algorithms for this setting that satisfy the safety constraint, detect changepoints, and restart without the knowledge of the number of changepoints or their locations. We provide regret bounds for our algorithms and show that the bounds are comparable to their counterparts from the safe bandit and piecewise i.i.d. bandit literature. We also provide the first matching lower bounds for this setting. Empirically, we show that our safety-aware algorithms perform similarly to the state-of-the-art actively adaptive algorithms that do not satisfy the safety constraint.

1 INTRODUCTION

Consider a startup XYZ that wants to maximize revenue collection from ad placements when users land on their webpage. Revenue is generated when users click on the ads. The user preferences change over time but not quickly enough so that XYZ can focus on maximizing the revenue collection for some time before changing its strategy again. To do this XYZ must detect the user's new preferences and modify its suggestions. Due to budget constraints XYZ must make sure that the aggregate revenue collection must not fall below a certain threshold. The difficulty is that XYZ does not know placing which ads will surely result in revenue above the threshold. This constrains XYZ from randomly placing different ads on their landing webpage. On consultation with the industry experts XYZ comes up with a default action that is known historically to be a highly favored by users. Hence, XYZ comes up with a new safety constraint such that when their algorithm is unsure of which ad to place for some user it can fall back on this default action. The learning algorithm now has to balance between exploration under safety constraints and exploitation in this slowly changing environment.

The dilemma faced by XYZ can be modeled as a sequential decision making problem in the piecewise i.i.d. bandit setting under safety constraints. In the piecewise i.i.d. bandit setting the learner is provided with a set of arms $i \in \{0, 1, 2, \dots, K\}$ where we index the default arm (baseline) as $i = 0$, and there exists a finite number of changepoints where the mean μ_i of one or more arms may change simultaneously. At every round $s \in \{1, 2, \dots, T\}$ the learner selects an action $I_s \in \{0, 1, 2, \dots, K\}$ and observes the feedback $X_{I_s}(s)$ where $\mathbb{E}[X_{I_s}(s)] =: \mu_{I_s}(s)$. Define i^* as the optimal arm such that $\mu_{i^*}(s) > \mu_i(s)$ for all i. The goal of the learner is to maximize reward by quickly finding the optimal arm i^* under the following safety constraint

$$\sum_{s=1}^{t} X_{I_s}(s) \geq (1-\alpha) \sum_{s=1}^{t} X_0(s) \tag{1}$$

for all $t \in \{1, \dots, T\}$, where $\alpha \in (0, 1]$ is the risk parameter, and T is the horizon. The constraint in eq. (1) represents how much the learner is allowed to risk in conducting the exploration. For example, if $\alpha \to 0$ the learner is expected to sample arms that are at least better than the baseline arm 0. The baseline arm represents expert's belief over the current user preferences and may change over time. Similar to the setting of Wu et al. [2016] we assume that the mean of the baseline arm is not known to the learner. Note that Wu et al. [2016] is not suited for the piecewise i.i.d setting. By change of belief of the expert we mean that only the value of the baseline arm changes and not the index.

The challenge in our setting is three-fold: **1)** Ensure that the safety constraint (1) on cumulative reward is satisfied.

Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022), PMLR 180:1402–1412.

Consider the scenario where the risk parameter $\alpha = 1$. In this case satisfying the eq. (1) is easy as choosing any arm satisfies the constraint. However as $\alpha \to 0$ maintaining the safety constraint becomes difficult as exploration becomes limited or it will violate the safety constraints. **2)** Adapt to the piecewise i.i.d. nature of the environment. Observe that as the means of arms change abruptly at changepoints the algorithm must adapt or the safety constraint will be violated. Further, to detect changepoints the algorithm must conduct additional exploration without violating the eq. (1). Note that Wu et al. [2016] do not consider any such piecewise i.i.d. setting. **3)** Finally, minimize the cumulative regret by quickly finding the optimal arm for each of the time segment between two changepoints. Our contributions are as follows:

1) We formulate the novel piecewise i.i.d. bandit setting under safety constraints. We show that the current state-of-the-art conservative algorithms [Wu et al., 2016] as well as the changepoint detection algorithms are not equipped to handle the safety constraint in eq. (1) in this setting.

2) We propose two actively adaptive algorithms that detect changepoints and restart by erasing past history of interactions. Simultaneously these algorithms ensure that the safety constraint is satisfied. The current changepoint detection algorithms [Besson and Kaufmann, 2019, Mukherjee and Maillard, 2019, Besson et al., 2020] do not take into account the safety constraint and hence are not suited for our setting.

3) We provide theoretical guarantees for both of our algorithms and uncover new problem dependent terms that depends on the optimality gaps, changepoint gaps, the gap of the baseline arm, and the risk parameter α. We also provide the first matching lower bounds for this setting. Empirically we show that our proposed methods perform comparably against safety oblivious changepoint detection algorithms.

2 RELATED WORKS

Our work lies at the intersection of two interesting areas: 1) Changepoint detection in piecewise i.i.d bandits, and 2) Safe Sequential Decision Making. In piecewise i.i.d bandits it is assumed that the change of mean (drift) of an arm are well separated and significant enough to be detected. The previous works in this setting are broadly classified into two groups, viz. passively adaptive and actively adaptive algorithms. Passively adaptive algorithms such as Discounted UCB (D-UCB) [Kocsis and Szepesvári, 2006], Sliding Window UCB (SW-UCB) [Garivier and Moulines, 2011], and Discounted Thompson Sampling (D-TS) [Raj and Kalyani, 2017] do not try to detect the changepoints and only focus on minimizing the regret over a short window of the time horizon. On the contrary, actively adaptive algorithms such as EXP3.R [Allesiardo et al., 2017], CD-UCB [Liu et al., 2017], CUSUM [Liu et al., 2017], M-UCB [Cao et al., 2018], GLR-UCB [Besson and Kaufmann, 2019, Besson et al., 2020], Ad-Switch [Auer et al., 2019], and UCB-CPD [Mukherjee and Maillard, 2019] try to detect the changepoints and restart by erasing all the past history of interactions. Actively adaptive algorithm like GLR-UCB, M-UCB, UCB-CPD has several advantages over passively adaptive algorithms. In environments where the changepoint gaps are large and well-separated the passively adaptive algorithms perform poorly (see [Besson and Kaufmann, 2019]). The EXP3.R (an adaptive version of EXP3.S [Auer et al., 2002b]) is more pessimistic than other actively adaptive algorithms like UCB-CPD, GLR-UCB as it uses the conservative exponential weighting algorithm EXP3 for changepoint detection, and hence, performs poorly in practice. GLR-UCB uses the Bernoulli generalized likelihood ratio test involving Kullback Leibler (KL) based divergence function as changepoint detector. The KL divergence function of GLR-UCB better exploits the geometry of (sub-)Bernoulli distributions and so it outperforms M-UCB, Ad-Switch . Note that none of the above algorithms are safety aware.

The safe sequential decision making setup has recently garnered a lot of attention in machine learning [Amodei et al., 2016, Turchetta et al., 2019]. Closer to our setting are the works that study regret minimization in bandits under safety constraints such as Wu et al. [2016], Kazerouni et al. [2017], Amani et al. [2019], Garcelon et al. [2020]. These works encode their safety requirements in the form of constraints on the cumulative rewards observed by the learner. This setup also called conservative bandits as the exploration is constrained by the constraints on the cumulative reward. Note that while Wu et al. [2016] studies the unstructured stochastic and adversarial bandit setting, the Kazerouni et al. [2017], Amani et al. [2019], Garcelon et al. [2020] study the linear bandit (structured) setting. Another line of related work [Moradipari et al., 2020, Pacchiano et al., 2020, Khezeli and Bitar, 2020] focuses on the idea of stagewise safety constraint where at every stage (round) the reward should be higher than a predetermined safety threshold with high probability. Note that our setting of safety constraints on cumulative rewards cannot be directly applied to the stagewise setting. None of the above works deal with the setting of piecewise i.i.d bandits with slow drift (change of means). Note that while Wu et al. [2016] studies an adversarial setting, they do not take any assumption on the reward distributions and hence their exploration scheme is highly conservative for piecewise i.i.d. bandits. Similarly, conservative bandits studying the contextual (structured) bandit setting [Kazerouni et al., 2017] do not use the known information about slow drift of means. Finally note that our setting is different than the thresholding bandit problem [Locatelli et al., 2016, Mukherjee et al., 2017] where the goal is to find all the arms above a fixed threshold B under the fixed budget setting.

Notations: Denote $[n] := \{1, 2, \ldots, n\}$. Define the set of

arms as $[K]$ indexed from $i = 1, 2, \ldots, K$. The baseline arm is denoted by the index 0. Note that the learner knows the index of the baseline arm but it does know the mean of the baseline arm. This is similar to the setting in Wu et al. [2016]. We define the set $[K]^+ \coloneqq [K] \cup \{0\}$ to indicate that baseline arm 0 is included as well. We define the mean of the arm i at round s as $\mu_i(s)$, and the empirical mean of the arm till round t as $\widehat{\mu}_i(t)$. We denote the optimal arm as i^* and the mean of the optimal arm at round s as $\mu_{i^*}(s)$. The reward of the arm i sampled at round s is denoted by $X_i(s)$. We assume that the rewards are coming from a bounded distribution supported on $[0, 1]$. We further denote the distribution of the i-th arm with mean $\mu_i(t)$ as $\nu(\mu_i(t))$. We denote the horizon (total rounds) as T. The safety threshold is denoted by B and the risk parameter is denoted by $\alpha \in [0, 1]$. For brevity we denote the sequence of rounds between s to t as $s : t$. For clarity of presentation we overload the notation $\mu_i(\cdot)$ to denote either the mean at round s as $\mu_i(s)$ or the mean over the rounds $1 : t$ as $\mu_i(1 : t)$. Similarly, $N(1 : t)$ denotes that number of pulls of i from $1 : t$. Define the empirical mean $\widehat{\mu}_i(1 : t) \coloneqq \frac{\sum_{s=1}^{t} X_{I_s} \mathbb{I}\{I_s = i\}}{\sum_{s=1}^{t} \mathbb{I}\{I_s = i\}} = \frac{\sum_{s=1}^{t} X_{I_s} \mathbb{I}\{I_s = i\}}{N_i(1:t)}$ where I_s denotes the arm pulled at round s.

3 GLOBAL CHANGEPOINT DETECTION

We now define the setup for the Global Changepoint Setting (GCS). Let the total number of changepoints till round T be denoted by G_T such that

$$G_T \coloneqq \#\left\{1 \leq s \leq T \mid \exists i \in [K]^+ : \mu_i(s-1) \neq \mu_i(s)\right\}. \tag{2}$$

We define the global changepoints $t_{c_0} < t_{c_1} < t_{c_2} < \ldots < t_{c_{G_T}}$ such that the g-th global changepoint is defined as:

$$t_{c_g} \coloneqq \inf\{s > t_{c_{g-1}} : \forall i \in [K]^+, \mu_i(s-1) \neq \mu_i(s)\}.$$

Hence at a global changepoint t_{c_g} the mean of all the arms $i \in [K]^+$ change simultaneously. Let $t_{c_0} = 1$ by convention. Note, that the baseline mean changes at changepoints to signify the new belief of the experts based on updated user preferences. We define the changepoint segment between t_{c_g} to $t_{c_{g+1}} - 1$ as ρ_g. Note that the optimal arm for each changepoint segment ρ_g may or may not be same. We further define a few notations. Let the confidence width of arm i for rounds $1 : t$ be defined as

$$\beta_i(1 : t, \delta) = \sqrt{\frac{2\log(4\log_2(t+1)/\delta)}{N_i(1:t)}} \tag{3}$$

with the standard condition that if $N_i(1 : t) = 0$ then $\beta_i(1 : t, \delta) = \infty$. In our case it suffices to take the leading constant of $\beta_i(1 : t, \delta)$ as 2, though tighter bounds are known and can be used in practice, e.g. Balsubramani [2014], Tanczos et al. [2017], Howard et al. [2021]. These type of anytime bounds constructed with $\beta_i(1 : t, \delta)$ are known to be tight in the sense that $\mathbb{P}\left(\bigcup_{t=1}^{\infty}\{|\widehat{\mu}_i(1:t) - \mu_i(1:t)| \geq \beta_i(1:t,\delta)\}\right) \leq \delta$ and that there exists an absolute constant $C \in (0, 1)$ such that $\mathbb{P}\left(\{|\widehat{\mu}_i(1:t) - \mu_i(1:t)| \geq C\beta_i(1:t,\delta) \text{ for infinitely many } t \in \mathbb{N}\}\right) = 1$ by the Law of the Iterated Logarithm [Hartman and Wintner, 1941]. Next we define the upper confidence bound for i as

$$U_i(1 : t) \coloneqq \widehat{\mu}_i(1 : t) + \beta_i(1 : t, \delta) \tag{4}$$

and the lower confidence bound from $1 : t$ as

$$L_i(1 : t) \coloneqq \widehat{\mu}_i(1 : t) - \beta_i(1 : t, \delta). \tag{5}$$

We define the UCB arm u_t at round t as

$$u_t \coloneqq \arg\max_{i \in [K]} U_i(1 : t) \tag{6}$$

which is the arm with the highest uncertainty and needs to be explored more to get a better estimate of its true mean [Agrawal, 1995, Auer et al., 2002a]. Finally, we define the empirical safety budget as

$$\widehat{Z}(1:t) \coloneqq \sum_{s=1}^{t-1} L_{I_s}(1:s) + L_{u_t}(1:t) - (1-\alpha)\sum_{s=1}^{t} U_0(1:t) \tag{7}$$

which quantifies by how much the safety constraint is being violated. Also recall that the baseline arm is indexed as 0, and the learner does not know the mean of the baseline arm. This is similar to the second setting in Wu et al. [2016].

3.1 SAFE GLOBAL RESTART ALGORITHM

In this section we introduce the Safe Global Restart (SGR) algorithm which is a safety aware global changepoint detection algorithm. SGR is an actively adaptive algorithm and so it restarts by erasing the history of interactions once it detects a changepoint. We define the parameter r_i for the i-th arm as the last restart round when a changepoint was detected and the arm i history was erased. We define the last restart vector

$$\boldsymbol{r} \coloneqq \{r_1, r_2, \ldots, r_K\} \cup \{r_0\}.$$

The safety budget for the GCS is $\widehat{Z}(1 : t)$

$$\begin{aligned} &\coloneqq \sum_{s=1}^{t-1} L_{I_s}(1 : s) + U_{u_t}(1 : t) - (1 - \alpha)\sum_{s=1}^{t} U_0(1 : t) \\ &\overset{(a)}{=} \sum_{s=1}^{t-1} L_{I_s}(r_{I_s} : s) + U_{u_t}(r_{u_t} : t) - (1 - \alpha)\sum_{s=1}^{t} U_0(r_0 : t) \end{aligned} \tag{8}$$

where, (a) follows because when a changepoint is detected the history is erased for that arm $i \in [K^+]$.

We now state the main aspects of SGR . SGR is initialized by sampling each arm once. Then at every round SGR decides

to pull the UCB arm u_t if $\widehat{Z}(1:t) \geq 0$ or the baseline arm 0 if $\widehat{Z}(1:t) < 0$. Then SGR samples the next arm, observes the reward $X_{I_t}(t)$ and updates the problem parameters. Finally SGR calls the CPD changepoint detector sub-routine to detect any changepoint. If a changepoint is detected at round t by CPD then it erases the history of interactions for all arms (including baseline arm) and sets the restarting time for all arms $i \in [K]^+$ as $r_i = t$. We state the pseudo-code of the policy SGR in Algorithm 1 and the key idea behind CPD in the following Section 3.2.

Algorithm 1 Safe Global Restart (SGR)

Input: Risk parameter $\alpha \in [0, 1)$
Set $r_i = 1, \forall i \in [K]^+$. Pull each arm once.
for $t = K^+ + 1, K^+ + 2, \ldots$ **do**
  if $\widehat{Z}(t) \geq 0$ **then**
    Set $I_t = u_t$ from eq. (6) ▷*Pull UCB arm*
  else if $\widehat{Z}(t) < 0$ **then**
    Set $I_t = 0$ ▷*Baseline arm*
  Pull I_t and observe $X_{I_t}(t)$.
  Update $\widehat{\mu}_{I_t}(r_{I_t} : t), N_{I_t}(r_{I_t} : t)$, and $\widehat{Z}(r_{I_t} : t)$ in eq. (7).
  Call CPD $(\boldsymbol{r}, t, \text{global})$ ▷*Call CPD*

3.2 CHANGEPOINT DETECTION

The sequential changepoint detection has a long history in the statistical community [Basseville et al., 1993, Wu, 2007]. We explain the sequential changepoint detection through the following example: Consider a single arm i. Let at some round t we have a collection of i.i.d. samples $X_i(1), X_i(2), \ldots, X_i(t)$ from a bounded distribution that is supported on $[0, 1]$. The goal of changepoint detection is to find out whether all the t samples have come from the same distribution with mean $\mu_i(1:t)$ or there exist a changepoint $\tau_{c_g} \in \mathbb{N}$ such that $X_i(1), X_i(2), \ldots, X_i(\tau_{c_g} - 1)$ have mean $\mu_i(1 : \tau_{c_g} - 1)$ while $X_i(\tau_{c_g}), X_i(\tau_{c_g} + 1), \ldots, X_i(t)$ have a different mean $\mu_i(\tau_{c_g} : t) \neq \mu_i(1 : \tau_{c_g} - 1)$. For notational convenience let us denote $\mu_i(1 : \tau_{c_g})$ as μ' and $\mu_i(\tau_{c_g} + 1 : t)$ as $\mu^{''}$ respectively. Hence, a sequential changepoint detector is defined as a stopping time $\tau^{chg} < \infty$ that rejects the null hypothesis $\mathcal{H}_0 : (\exists \mu' : \forall i \in \mathbb{N}, \mathbb{E}[X_i] = \mu')$ in favor of the alternate hypothesis $\mathcal{H}_1 : (\exists \mu^{''} \neq \mu', \tau_{c_g} \in \mathbb{N} : X_i(1), X_i(2), \ldots, X_i(\tau_{c_g}) \sim \nu(\mu'),$ and $X_i(\tau_{c_g}+1), X_i(\tau_{c_g}+2), \ldots, X_i(t) \sim \nu(\mu^{''}))$. Previous works have studied the Generalized Likelihood Ratio Test (GLRT) to detect the changepoints using this hypothesis testing idea. Many of the previous works [Wilks, 1938, Siegmund and Venkatraman, 1995, Maillard, 2019, Besson and Kaufmann, 2019, Besson et al., 2020] that have studied this setting used Generalized Likelihood Ratio Test (GLRT) to detect changepoints. The GLRT test works as follows: We first calculate the GLRT statistic defined by

$$\log \frac{\sup_{\mu', \mu'', \tau_{c_g} < t} L\left(X_i(1), \ldots, X_i(t); \mu', \mu^{''}, \tau_{c_g}\right)}{\sup_{\mu'} L\left(X_i(1), \ldots, X_i(t); \mu'\right)}$$

where we denote the term $L(X_i(1), \ldots, X_i(t); \mu')$ and $L(X_i(1), \ldots, X_i(t); \mu', \mu^{''}, \tau)$ as the likelihood of the first t observations under hypothesis $\mathcal{H}_0$ and $\mathcal{H}_1$ respectively. Now if the GLRT statistic crosses a threshold $\widetilde{\beta}(t, \delta)$ then it indicates that there exists a changepoint and the null hypothesis $\mathcal{H}_0$ is rejected. A similar type of test, called the CUSUM test [Page, 1954, Liu et al., 2017] has also been studied where the distributions $\nu(\mu_1)$ and $\nu(\mu_2)$ are completely known. Note that GLRT works in the case when both distributions are unknown but they come from the same canonical exponential family. A detailed discussion on this can be found in Maillard [2019].

Confidence-based scan statistic: An alternative to the GLRT based scan statistics is the confidence-based scan statistic that have been studied in Mukherjee and Maillard [2019]. In the confidence based scan statistic the total number of samples of arm i from $1 : t$ is divided into slices, and for each slice s a confidence interval is built of the form

$$\widehat{\mu}_i(1:s) \pm \beta_i(1:s, \delta) \text{ and } \widehat{\mu}_i(s+1:t) \pm \beta_i(s+1:t, \delta) \tag{9}$$

where $\beta_i(1 : s, \delta)$ is from eq. (3). Now if there exists some s at which the confidence intervals do *not* overlap such that

$$\tau_{c_g} := \inf \Big\{ t \in \mathbb{N} : \exists i \in [K]^+, \exists s \in [1, t], |\widehat{\mu}_i(1:s) - \widehat{\mu}_i(s+1:t)| > \beta_i(1:s, \delta) + \beta_i(s+1:t, \delta) \Big\} \tag{10}$$

then report a changepoint at s. Besson et al. [2020] show that GLRT outperforms the confidence-based scan statistic as it better exploits the geometry of the Bernoulli distributions. In our work we use the confidence-based scan statistic as our goal is not to compete in the vanilla changepoint detection setting but to derive novel bounds for the safety aware piecewise i.i.d. setup proposed in this work.

Finally, we propose the CPD changepoint detector sub-routine which is similar to the UCB-CPD algorithm in Mukherjee and Maillard [2019]. The CPD takes input the restart vector $\boldsymbol{r}$, current time t, and the type $\in \{\text{global}, \text{local}\}$ indicating whether it is a global or a local changepoint setting. In this section we only discuss the global setting while the local setting is discussed in Section 5. The CPD divides the total rounds $r_i : t$ into $(t - r_i)$ slices for each arm $i \in [K]^+$ and then proceeds to conduct the confidence-based scan statistics as discussed in (10). If there is a disjoint slice s then CPD reports a changepoint at s, then erases the history of interactions (including the baseline arm) and resets the restarting round counter $r_i, \forall i \in [K]^+$

to the current time t. We also reset the safe budget $\widehat{Z}(1:t)$ to 0. Ideally in practice we can still continue the accrued safe budget to the next changepoint section from $\tau_{c_{g+1}}+1$ without setting it to 0, but this makes our theoretical analysis more tedious.

Algorithm 2 CPD $(\boldsymbol{r}, t, \text{type})$

for $i = 1, 2, \ldots, K^+$ **do**
  for $t' = r_i, r_i+1, \ldots, t$ **do**
    if $L_i(r_i:t') < U_i(t'+1:t)$ or $U_i(r_i:t') > L_i(t'+1:t)$ **then**
      if type = global **then**
        $\{\widehat{\mu}_j(r_i:s), N_j(r_i:s)\}_{s=r_j}^{t} = \{0,0\}_{s=r_j}^{t}$, $\forall j \in [K^+]$. Set $r_j = t, \forall j \in [K^+]$. Set $\widehat{Z}(1:t) = 0$.
      else
        $\{\widehat{\mu}_i(r_i:s), N_i(r_i:s)\}_{s=r_i}^{t} = \{0,0\}_{s=r_i}^{t}$, $r_i = t$. Set $\widehat{Z}(1:t) = 0$.

3.3 REGRET ANALYSIS FOR SGR

We denote the time interval segment $\rho_g := [t_{c_g}, t_{c_{g+1}}-1]$ so that the segment ρ_g starts at round t_{c_g} and ends at $t_{c_{g+1}-1}$. Let $\mu_{i,g}$ denote the mean of arm i for the segment ρ_g. Let the changepoint gap between the segments ρ_g and ρ_{g+1} be $\Delta_{i,g}^{chg} := |\mu_{i,g} - \mu_{i,g+1}|$. We redefine the optimality gap for the segment ρ_g as $\Delta_{i,g}^{opt} := \mu_{i^*,g} - \mu_{i,g}$. Let τ_{c_g} denote the first round when the changepoint t_{c_g} is detected and SGR is restarted. Then we define the quantity $N_{0,g}^{chg}$ as the number of times the baseline arm is sampled from rounds t_{c_g} till the detection of changepoint at τ_{c_g}. Finally, we define the delay of detection of the g-th changepoint as

$$d_g := \left\lceil K + \left(\max_{i\in[K]} \frac{B(T,\delta)}{(\Delta_{i,g}^{chg})^2} + \frac{B(T,\delta)}{(\Delta_{0,g}^{chg})^2} + N_{0,g}^{bse} \right) 4K \right\rceil \quad (11)$$

such that SGR detects the change at t_{c_g} within $t_{c_g}+1$ till $t_{c_g}+d_g$ rounds with probability greater than $1-\delta$. We define the quantity $B(T,\delta) = 16\log(4\log_2(T/\delta))$. The quantity $N_{0,g}^{chg}$ denotes the number of samples of the baseline arms after the changepoint t_{c_g} has occurred but not detected and is defined by

$$N_{0,g}^{bse} := \frac{1}{\alpha\mu_{0,g}} \sum_{i\in[K]} \frac{B(T,\delta)}{\max\{\Delta_{i,g}^{opt}, \Delta_{0,g}^{opt} - \Delta_{i,g}^{opt}\}}.$$

Intuitively, $N_{0,g}^{chg}$ is the number of samples required after t_{c_g} has occurred and the safe budget $\widehat{Z}(1:t)$ falls below 0. In $N_{0,g}^{chg}$ if α is very small, we can still explore other arms as long as the baseline arm 0 is close to the optimal arm $\mu_{i,g}^*$ (so that $\Delta_{0,g}^{opt}$ is small) while the other arms are clearly sub-optimal (i.e. the $\Delta_{i,g}^{opt}$ are large). If this happens then the sub-optimal arms are quickly discarded, while the $\widehat{Z}(1:t)$ stays positive and the regret penalty is small. We now define a mild assumption on separation of changepoints which is standard in changepoint detection settings (see Besson and Kaufmann [2019], Besson et al. [2020]). Without this assumption the changepoints can be too frequent and cannot be detected before the next change happens. We require this assumption for our theoretical guarantees. Note that in the experiments we show that even when this assumption does not hold our proposed algorithms performs well.

Assumption 1. ***(Separation of changepoints for GCS)*** *We assume that the for all $g \in \{0,1,2,\ldots,G_T\}$ two consecutive changepoints t_{c_g} and $t_{c_{g+1}}$ are separated as $t_{c_{g+1}} - t_{c_g} \geq 2\max\{d_g, d_{g+1}\}$, where d_g is stated in* (11).

The Assumption 1 assumes that two consecutive changepoints are separated enough to be detected by the changepoint detector. Note that our detection delay d_g is larger than Besson et al. [2020] because between $t_{c_g} : \tau_{c_g}$ the budget $\widehat{Z}(1:t)$ may fall below 0 and SGR may need to sample the baseline arm from the next segment ρ_{g+1}. We denote an event by ξ and its complement by $\overline{\xi}$. Define the good event ξ_g^{del} that all changepoints $g' \leq g$ have been detected with delay at most $d_{g'}$. Let the safe budget time set $\mathcal{Q}(1:t) := \left\{ s \in [1:t] : \widehat{Z}(1:s) \geq 0 \right\}$ be the set of all rounds $1:t$ when $\widehat{Z}(1:s) \geq 0$. We can decompose the expected regret as

$$\sum_{g=1}^{G_T} \Bigg[\underbrace{\sum_{i=1}^{K} \sum_{s\in\mathcal{Q}(\tau_{c_{g-1}}:t_{c_g}-1)} \Delta_i^{opt}(s)\mathbb{E}[N_i(s)|\xi_g^{del}(s)]\mathbb{P}\left(\xi_g^{del}(s)\right)}_{\textbf{Part (A), UCB arm pulled, Safe budget } \widehat{Z}(\tau_{c_{g-1}}:s)\geq 0}$$
$$+ \underbrace{\sum_{s\in\overline{\mathcal{Q}}(\tau_{c_{g-1}}:t_{c_g}-1)} \Delta_0^{opt}(s)\mathbb{E}[N_0(s)|\xi_g^{del}(s)]\mathbb{P}\left(\xi_g^{del}(s)\right)}_{\textbf{Part (B), Baseline arm pulled, Safe budget } \widehat{Z}(\tau_{c_{g-1}}:s)<0}$$
$$+ \underbrace{\sum_{i=1}^{K} \sum_{s\in\mathcal{Q}(t_{c_g}:\tau_{c_g}-1)} \Delta_i^{opt}(s)\mathbb{E}[N_i(s)|\xi_g^{del}(s)]\mathbb{P}\left(\xi_g^{del}(s)\right)}_{\textbf{Part (C), Changepoint Pulls, Safe budget } \widehat{Z}(\tau_{c_{g-1}}:s)\geq 0}$$
$$+ \underbrace{\sum_{s\in\overline{\mathcal{Q}}(t_{c_g}:\tau_{c_g}-1)} \Delta_0^{opt}(s)\mathbb{E}[N_0(s)|\xi_g^{del}(s)]\mathbb{P}\left(\xi_g^{del}(s)\right)}_{\textbf{Part (D), Changepoint Baseline Pulls, Safe budget } \widehat{Z}(\tau_{c_{g-1}}:s)<0}$$
$$+ \sum_{s=\tau_{c_{g-1}}}^{T} \underbrace{\mathbb{P}(\overline{\xi_g^{del}(s)})}_{\textbf{Part (E), Total Detection Delay Error}} \Bigg], \quad (12)$$

which follows by dividing the total rounds till T into G_T segments when the changepoint t_{c_g} is detected at τ_{c_g}. We then further subdivide it into two parts $\tau_{c_{g-1}} : t_{c_g}-1$ (rounds before t_{c_g}) and $t_{c_g} : \tau_{c_g}-1$ (rounds before detection of t_{c_g}. The four parts (A)-(D) further divides the

two time segments $\tau_{c_{g-1}} : t_{c_g} - 1$ and $t_{c_g} : \tau_{c_g} - 1$ based on the available safe budget and using the definition of $\mathcal{Q}(1:t)$. Now using Assumption 1 we can show that two consecutive changepoints are separated enough to correctly control the pulls of the baseline arm, and detect the optimal arm given that ξ_g^{del} holds. The main difference from previous changepoint detection works like Besson and Kaufmann [2019], Mukherjee and Maillard [2019] lies in controlling the detection delay and false alarm under the safety budget constraint. Using the Assumption 1 and changepoint detection Lemma 3 we can show that the detection delay is bounded by $2\max\{d_g, d_{g+1}\}$ with high probability. We now define a few problem dependent parameters which is key to analyze the regret of SGR. We define the quantity $H_{i,g}^{(1)} := \max\left\{\frac{1}{\Delta_{i,g-1}^{opt}}, \frac{\Delta_{i,g-1}^{opt}}{\left(\Delta_{i,g-1}^{chg}\right)^2}\right\}$ as the hardness of discarding the sub-optimal arm i and avoiding false detection, the quantity $H_{i,g}^{(2)} := \max_{j\in[K]^+} \frac{\Delta_{i,g}^{opt}}{\left(\Delta_{j,g}^{chg}\right)^2}$ as the hardness for detecting the changepoint t_{c_g} due to i after the changepoint has happened. Finally, the quantity $H_{i,g}^{(3)} := \frac{\Delta_{\max,g}^{opt}}{\max\{\Delta_{i,g}^{opt}, \Delta_{0,g}^{opt} - \Delta_{i,g}^{opt}\}}$ captures the trade-off of selecting the baseline arm 0 once the changepoint t_{c_g} occurred. The regret of SGR is shown below.

Theorem 2. *Let $H_{i,g}^{(1)}, H_{i,g}^{(2)}, H_{i,g}^{(3)}$ is defined above for the segment ρ_g. Then the expected regret of SGR is upper bounded by*

$$\mathbb{E}[R_T] \leq O\Bigg(\Bigg(\sum_{g=1}^{G_T}\sum_{i=1}^{K^+}\left(H_{i,g-1}^{(1)} + H_{i,g}^{(2)}\right) + \sum_{g=1}^{G_T}\frac{1}{\alpha\mu_{0,g-1}}\sum_{i=1}^{K}H_{i,g-1}^{(3)} + K\sum_{g=1}^{G_T}\frac{1}{\alpha\mu_{0,g}}\sum_{i=1}^{K}H_{i,g}^{(3)}\Bigg)\log\left(\frac{\log_2 T}{\delta}\right)\Bigg). \tag{13}$$

In the result of (13) the first term is the optimality regret suffered before discarding the arm i when the safety budget $\widehat{Z}(1:t) \geq 0$. The second term denotes the regret suffered for the changepoint detection due to arm i. The third term is the regret suffered for the section ρ_{g-1} when the safety budget $\widehat{Z}(1:t) < 0$. Finally, the fourth term is the regret suffered due to the changepoint t_{c_g} and safety budget $\widehat{Z}(1:t) < 0$. SGR conducts no forced exploration which results in a fully gap-dependent bound. This result is different than the gap-dependent bound in Mukherjee and Maillard [2019] which does not contain the third and fourth terms in (13). The bound in (13) is more informative than Corollary 1 as it correctly captures the dependence with respect to gaps for each segment ρ_g.

4 LOCAL CHANGEPOINT DETECTION

In the Local Changepoint Setting (LCS) at any changepoint at least one arm has a change of mean. Recall G_T from (2). We then define the local changepoints $t_{c_0} < t_{c_1} < \ldots < t_{c_{G_T}}$ such that the g-th local changepoint is defined as

$$t_{c_g} := \inf\{s > t_{c_{g-1}} : \exists i \in [K], \mu_i(s-1) \neq \mu_i(s)\}.$$

So at a local changepoint the mean of one or more arms may change simultaneously. Let $G_T^i := \#\{1 \leq s \leq T \mid \mu_i(s-1) \neq \mu_i(s)\}$ denote the number of changepoints only for the i-th arm. It follows that $G_T^i \leq G_T$ but for some arms there could be arbitrary difference between these two quantities. Note that $J_T := \sum_{i=1}^K G_T^i \leq KG_T$. Define $t_{c_g}^i := \inf\{s > t_{c_{g-1}}^i : \mu_i(s-1) \neq \mu_i(s)\}$ as the g-th changepoint for the i-th arm. Again $t_{c_0}^i = 1$ for all arms $i \in [K]^+$ by convention. We denote the segment between rounds $t_{c_g}^i$ and $t_{c_{g+1}}^i - 1$ as ρ_g^i.

Consider the scenario that a learner has figured out the best arm i in a segment ρ_g^i but then at a local changepoint $t_{c_g}^j$, an arm $j \neq i$ becomes the new optimal arm (but arm i does not change). So it will not be able to detect $t_{c_g}^j$ and will continue sampling arm i. Hence the leaner need to conduct forced exploration of all arms to have a good estimate of all arms. This idea is shown in Algorithm 3 where at every round SLR first checks that the safety budget is positive and then either conducts forced exploration of arms with exploration parameters γ (to be defined later) or samples the UCB arm u_t. If the safety budget is negative then SLR samples the baseline arm so that the budget becomes positive and SLR can explore again. Finally SLR calls the CPD sub-routine with type as "local" to detect local changepoints for an arm. Once the changepoint is detected it restarts only that arm as this is the local changepoint setting. The crucial thing to note is that we conduct forced exploration only when safety budget is available (positive).

Algorithm 3 Safe Local Restart (SLR)

Input: Risk parameter α, exploration factor γ
Set $r_i = 1, \forall i \in [K^+]$. Pull each arm once.
for $t = K^+ + 1, K^+ + 2, \ldots$ **do**
  if $\widehat{Z}(t) \geq 0$ **then**
    if $t \mod \lfloor \frac{K}{\gamma} \rfloor \notin [K]$ **then**
      Set $I_t = u_t$ from eq. (6) ▷*Pull UCB arm*
    else if $t \mod \lfloor \frac{K}{\gamma} \rfloor \in [K]$ **then**
      Set $I_t = t \mod \lfloor \frac{K}{\gamma} \rfloor$ ▷*Forced Exploration*
  else if $\widehat{Z}(t) < 0$ **then**
    Set $I_t = 0$ ▷*Baseline arm*
  Pull I_t and observe $X_{I_t}(t)$.
  Update $\widehat{\mu}_{I_t}(r_{I_t} : t), N_{I_t}(r_{I_t} : t)$,and $\widehat{Z}(r_{I_t} : t)$ in eq. (7).
  Call CPD $(\boldsymbol{r}, t, \text{local})$ ▷*Call CPD*

4.1 REGRET ANALYSIS FOR SLR

We can extend the analysis of SGR to also bound the regret for SLR. The key difference between the two analysis

is that SLR needs to bound the regret for each segment ρ_g^i for all arms $i \in K^+$. To this effect we first redefine changepoint gap for any arm i between the segments ρ_g^i and ρ_{g+1}^i as $\Delta_{i,g}^{chg} := |\mu_{i,g} - \mu_{i,g+1}|$, and the optimality gap as $\Delta_{i,g}^{opt} := \mu_{i^*,g} - \mu_{i,g}$. Let $\tau_{c_g}^i$ denote the first round when the changepoint $t_{c_g}^i$ is detected and SLR is restarted for the arm i. Then we define detection delay for the changepoint at $t_{c_g}^i$ as

$$d_{i,g} := \left\lceil \frac{K}{\gamma} + \frac{4}{\gamma}\left(\frac{B(T,\delta)}{(\Delta_{i,g}^{chg})^2} + \frac{B(T,\delta)}{(\Delta_{0,g}^{chg})^2} + N_{0,g}^{bse} \right) \right\rceil. \quad (14)$$

such that $t_{c_g}^i$ is detected within $t_{c_g}^i + 1 : t_{c_g}^i + d_{i,g}$ rounds and γ is the exploration rate of SLR . Again we denote $B(T,\delta) = 16\log(4\log_2(T/\delta))$. Note that the delay $d_{i,g}$ scales with the exploration rate γ so that SLR while conducting forced exploration can detect $t_{c_g}^i$. Similar assumption has also been taken in Besson and Kaufmann [2019], Besson et al. [2020]. We then define the following assumption for the separation of changepoints between $t_{c_g}^i$ and $t_{c_{g+1}}^i$. Again note that this assumption is only required for theoretical guarantees. Empirically we show that SLR performs well even when the Assumption 2 is violated.

Assumption 2. ***(Separation of changepoints for LCS)*** *We assume that the for all $g \in \{0,1,2,\dots,G_T^i\}$ two consecutive changepoints $t_{c_g}^i$ and $t_{c_{g+1}}^i$ are separated as $t_{c_{g+1}}^i - t_{c_g}^i \geq 2\max\{d_{i,g}, d_{i,g+1}\}$, where $d_{i,g}$ is defined in* (14).

Next we introduce the quantity $\overline{H_{i,g}^{(2)}} := \frac{\Delta_{i,g}^{opt}}{(\Delta_{i,g}^{chg})^2}$ as the hardness for detecting the g-th changepoint for the arm i. Note that in the LCS setting the SLR algorithm is restarted only for the arm i and so in the hardness we do not see the max over all arms like the SGR setting. Finally using the Assumption 2 and the same analysis as in Theorem 2 but for each segment ρ_g^i and each arm $i \in [K^+]$ we bound the regret for SLR in Theorem 3.

Theorem 3. *Let $H_{i,g}^{(1)}, \overline{H_{i,g}^{(2)}}, H_{i,g}^{(3)}$ is defined above for the segment ρ_g. Then the expected regret of SLR is bounded by*

$$\mathbb{E}[R_T] \leq O\Bigg(\Bigg(\sum_{i=1}^{K^+}\sum_{g=1}^{G_T^i}\Big(H_{i,g-1}^{(1)} + \overline{H_{i,g}^{(2)}}\Big) + \sum_{i=1}^{K}\sum_{g=1}^{G_T^i}\frac{1}{\alpha\mu_{0,g-1}}\sum_{i=1}^{K}H_{i,g-1}^{(3)} + K\sum_{g=1}^{G_T}\frac{1}{\alpha\mu_{0,g}}\sum_{i=1}^{K}H_{i,g}^{(3)}\Bigg)\log\left(\frac{\log_2 T}{\delta}\right)\Bigg) + \gamma T. \quad (15)$$

The result in (15) has a similar interpretation to (13) (but with respect to each arm segment ρ_g^i instead of global segment ρ_g) except the gap-independent term of γT which results from the forced exploration of arms. We state the following corollary to summarize the result of SGR and SLR in the "easy" case when all the gaps are same.

Corollary 1. ***(Gap independent bound)*** *Setting $\Delta_{i,g}^{opt} = \Delta_{i,g}^{chg} = \sqrt{\frac{K\log T}{T}}$ for all $i \in [K]^+$ and exploration rate $\gamma = \sqrt{\frac{\log T}{T}}$ we obtain the gap independent regret upper bound of SGR and SLR as*

$$\mathbb{E}[R_T] \leq O\left(G_T K\sqrt{KT\log T} + \frac{G_T\log T}{\alpha\mu_{0,\min}}\right), \textbf{(SGR)}$$

$$\mathbb{E}[R_T] \leq O\left(G_T\sqrt{KT\log T} + \frac{G_T\log T}{\alpha\mu_{0,\min}}\right), \textbf{(SLR)}$$

where α is the risk parameter.

Comparing the above result with GLR-UCB (see Proposition 4) we see that SGR (or SLR) picks up an additional factor of $1/(\mu_{0,min}\alpha)$ per changepoint which signifies the hardness of finding the safe set of actions for maintaining the safety constraint (1). Further note that SGR suffers an extra factor of K in its bound compared to SLR . This is because in the GCS setting the algorithm restarts by erasing the history of interactions for all arms. Hence, our result mirrors a similar observation in Besson et al. [2020]. Moreover as $\alpha \to 0$ (risky setting) the regret increases proportionally. This is similar to the gap-independent bound in Wu et al. [2016] shown in Proposition 2 which holds for the stochastic setting without any changepoints. The key takeaway from this result is that the piecewise i.i.d. setting under safety constraints is no harder than the conservative stochastic setting of Wu et al. [2016] and piecewise i.i.d. setting given the changepoints are sufficiently separated. Finally we state the lower bound in the safe GCS setting.

Theorem 3. ***(Lower Bound)*** *Let $\mathcal{E}, \overline{\mathcal{E}}$ be two bandit environment and there exits a global changepoint at $t_{c_1} = T/2$. Let $\alpha > 0$ be the safety parameter and $\mu_{0,\min}$ be the mean of the minimum safety mean over the changepoint segments. Then the lower bound is given by $\mathbb{E}_{\mathcal{E},\overline{\mathcal{E}}}[R_T] \geq \left\{\frac{K}{(16e+8)\alpha\mu_{0,\min}} + \frac{\log T}{\alpha\mu_{0,\min}}, \frac{\sqrt{KT}}{\sqrt{32e+16}} + \frac{\log T}{\alpha\mu_{0,\min}}\right\}$.*

The proof is given in Appendix .6 and follows from the change of measure argument. Additionally, we use the lower bound results from safe bandit setting of Wu et al. [2016] and changepoint detection setting of Gopalan et al. [2021] to arrive at the final result. Note that both of these works do not take into account the safe GCS setting. Finally, comparing the results of Theorem 3 and Corollary 1 we see that SGR matches the lower bound when $G_T = 1$ except a factor of $O(K\sqrt{\log T})$. Similarly, since GCS is a special case of LCS , we see that SLR also matches the lower bound except a factor of $O(K\sqrt{\log T})$.

5 EXPERIMENTS

In this section we test SGR and SLR against safety oblivious actively adaptive algorithms GLR-UCB , UCB-CPD as well

as passive algorithm D-UCB , and safety aware algorithms CUCB , and UMOSS . A detailed discussion on the algorithms, hyper-parameter tuning, and time complexity of the algorithms is given in Appendix .7. One further experiment showing the performance of SGR , SLR under different values of α is shown in Appendix .7. All codes are provided in supplementary material.

Global Changepoint: In this setting all the arms (including baseline) change at every changepoint. The environment consist of 6 arms (including baseline) and the evolution of means with respect to rounds is shown in Figure 1 (Left). The three global changepoints are at $t = 2000, 4000$ and 6000. We set risk parameter $\alpha = 0.7$. The performance of all the algorithms is shown in Figure 2 (Left). The adaptive algorithms like UCB-CPD , GLR-UCB perform well as they detecting the changepoints and restart but they do not satisfy the safety constraints. Note that SGR performs similar to GLR-UCB , UCB-CPD as it also detects the changepoints and restarts as well as satisfy the safety constraint. It outperforms passive algorithm D-UCB , and safety aware algorithm CUCB . The safety aware algorithm CUCB is not suited for the safety constraint (1) under piecewise i.i.d. setting as it always chooses the baseline arm and fail to achieve sub-linear regret.

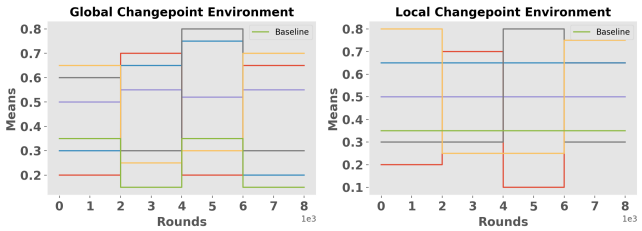

Figure 1: (Left) Global changepoint environment with $T = 8000$, $K^+ = 6$ and changepoints at $t = 2000, 4000$ and 6000. (Middle) Local changepoint environment with $T = 8000$, $K^+ = 6$ and changepoints at $t = 2000, 4000$ and 6000. Note that some arms do not change at these changepoints.

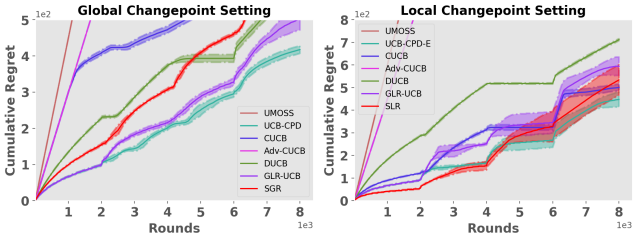

Figure 2: (Left) GCS setting with 3 changepoints and 6 arms. (Right) LCS setting with 3 changepoints and 6 arms.

Local Changepoint: In this setting at least one arm changes at every changepoint. We show that the environment in Figure 1 (Right) and the performance of all the algorithms in Figure 2 (Right). The three local changepoints are at $t = 2000, 4000$ and 6000. We set a constant baseline $\mu_0 = 0.35$, and risk parameter $\alpha = 0.7$. Again we see that the safety aware algorithm CUCB fail to achieve sub-linear regret as it always chooses the baseline arm. On the other hand adaptive algorithms like UCB-CPDE , GLR-UCB performs well in detecting the changepoints but they do not satisfy the safety constraints. Note that SGR performs similar to GLR-UCB , UCB-CPD as it also detects the changepoints and restarts as well as satisfy the safety budget. It outperforms passive algorithm D-UCB , and safety aware algorithm CUCB .

Real Setting: We show a real world experiment on the Movielens Dataset. In this experiment none of our modeling assumptions hold. We experiment with the Movielens dataset from February 2003 [Harper and Konstan, 2016], where there are 6k users who give 1M ratings to 4k movies. We obtain a rank-4 approximation of the dataset over 128 users and 128 movies such that all users prefer either movies 7, 13, 16, or 20 (4 user groups). The movies are the arms and we choose 30 movies that have been rated by all the users. Hence, this testbed consists of 30 arms and is run over $T = 8000$. The changepoints are at $t = 2000$, $t = 4000$, and $t = 6000$. Note that at each changepoint the means of some arms may or may not change so this is LCS . For every changepoint segment, we uniform randomly sample an user from different user groups to simulate the piecewise i.i.d environment such that there is a change in the optimal arm. In this environment each arm has has a Gaussian distribution associated with it, where its mean evolve as shown in Figure 1 (Left). The baseline arm is set as 0.35. As shown in Figure 1 (Right), in this environment SLR outperforms all the other algorithms including CUCB and Adv-CUCB . This is because the means of the arms are close to each other and the baseline arm mean is close to them.

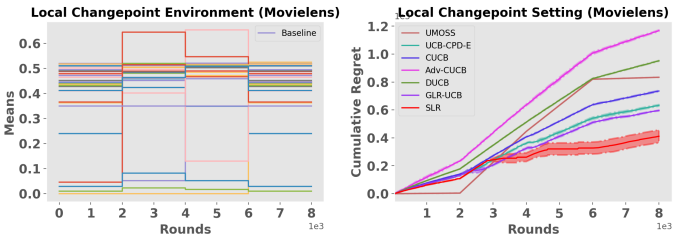

Figure 3: (Left) LCS setting with 3 changepoints and 30 arms. (Right) Regret in Movielens dataset.

6 CONCLUSION AND FUTURE WORKS

In this paper we studied the safety aware piecewise i.i.d. bandits under a new safety constraint. We proposed two actively adaptive algorithms SGR and SLR which satisfy the safety constraints as well as detect changepoints and restart. We provided regret bounds on our algorithms and showed how the bounds compare with respect to safety aware bandits

as well as adaptive algorithms. We also provided the first matching lower bounds for this setting. Future works include extending our setting to the rested and sleeping bandit setting under safety constraints. We also intend to explore experimental design approaches to piecewise i.i.d settings as in Pukelsheim [2006], Mason et al. [2021], Mukherjee et al. [2022]. Finally, incorporating variance aware techniques [Audibert et al., 2009, Mukherjee et al., 2018] may further improve the the performance of our proposed algorithms.

References

Rajeev Agrawal. 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, 1995.

Robin Allesiardo, Raphaël Féraud, and Odalric-Ambrym Maillard. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 3(4):267–283, 2017. doi: 10.1007/s41060-017-0050-5. URL https://doi.org/10.1007/s41060-017-0050-5.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9252–9262, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/09a8a8976abcdfdee15128b4cc02f33a-Abstract.html.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009. doi: 10.1016/j.tcs.2009.01.016. URL https://doi.org/10.1016/j.tcs.2009.01.016.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b. doi: 10.1137/S0097539701398375. URL https://doi.org/10.1137/S0097539701398375.

Peter Auer, Yifang Chen, Pratik Gajane, Chung-Wei Lee, Haipeng Luo, Ronald Ortner, and Chen-Yu Wei. Achieving optimal dynamic regret for non-stationary bandits without prior information. In *Conference on Learning Theory*, pages 159–163. PMLR, 2019.

Akshay Balsubramani. Sharp finite-time iterated-logarithm martingale concentration. *arXiv preprint arXiv:1405.2639*, 2014.

Michele Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice hall Englewood Cliffs, 1993.

Lilian Besson and Emilie Kaufmann. The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits. *arXiv preprint arXiv:1902.01575*, 2019.

Lilian Besson, Emilie Kaufmann, Odalric-Ambrym Maillard, and Julien Seznec. Efficient change-point detection for tackling piecewise-stationary bandits. 2020.

Yang Cao, Wen Zheng, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure for piecewise-stationary bandit: a change-point detection approach. *arXiv preprint arXiv:1802.03692*, 2018.

Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Improved algorithms for conservative exploration in bandits. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3962–3969. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/5812.

Aurélien Garivier and Eric Moulines. 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, 2011. doi: 10.1007/978-3-642-24412-4_16. URL https://doi.org/10.1007/978-3-642-24412-4_16.

Aditya Gopalan, Braghadeesh Lakshminarayanan, and Venkatesh Saligrama. Bandit quickest changepoint detection. *Advances in Neural Information Processing Systems*, 34, 2021.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):19, 2016.

Philip Hartman and Aurel Wintner. On the law of the iterated logarithm. *American Journal of Mathematics*, 63(1):169–176, 1941.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49 (2):1055–1080, 2021.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi, and Benjamin Van Roy. Conservative contextual linear bandits. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3910–3919, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/bdc4626aa1d1df8e14d80d345b2a442d-Abstract.html.

Kia Khezeli and Eilyan Bitar. Safe linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10202–10209, 2020.

Levente Kocsis and Csaba Szepesvári. Discounted ucb. *2nd PASCAL Challenges Workshop*, pages 784–791, 2006.

Fang Liu, Joohyun Lee, and Ness B. Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. *CoRR*, abs/1711.03539, 2017. URL http://arxiv.org/abs/1711.03539.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, NY, USA, June 19-24, 2016*, 48:1690–1698, 2016. URL http://jmlr.org/proceedings/papers/v48/locatelli16.html.

Odalric-Ambrym Maillard. Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds. In *Algorithmic Learning Theory*, pages 610–632, 2019.

Blake Mason, Romain Camilleri, Subhojyoti Mukherjee, Kevin Jamieson, Robert Nowak, and Lalit Jain. Nearly optimal algorithms for level set estimation. *arXiv preprint arXiv:2111.01768*, 2021.

Ahmadreza Moradipari, Christos Thrampoulidis, and Mahnoosh Alizadeh. Stage-wise conservative linear bandits. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/804741413d7fe0e515b19a7ffc7b3027-Abstract.html.

Subhojyoti Mukherjee and Odalric-Ambrym Maillard. Distribution-dependent and time-uniform bounds for piecewise iid bandits. *ICML Workshop on Reinforcement Learning for Real Life, 2019*, 2019.

Subhojyoti Mukherjee, Kolar Purushothama Naveen, Nandan Sudarsanam, and Balaraman Ravindran. Thresholding bandits with augmented UCB. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2515–2521. ijcai.org, 2017. doi: 10.24963/ijcai.2017/350. URL https://doi.org/10.24963/ijcai.2017/350.

Subhojyoti Mukherjee, KP Naveen, Nandan Sudarsanam, and Balaraman Ravindran. Efficient-ucbv: An almost optimal algorithm using variance estimates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Subhojyoti Mukherjee, Ardhendu S Tripathy, and Robert Nowak. Chernoff sampling for active testing and extension to active regression. In *International Conference on Artificial Intelligence and Statistics*, pages 7384–7432. PMLR, 2022.

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter L. Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. *CoRR*, abs/2006.10185, 2020. URL https://arxiv.org/abs/2006.10185.

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

Vishnu Raj and Sheetal Kalyani. Taming non-stationary bandits: A bayesian approach. *CoRR*, abs/1707.09727, 2017. URL http://arxiv.org/abs/1707.09727.

David Siegmund and ES Venkatraman. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics*, pages 255–271, 1995.

Ervin Tanczos, Robert Nowak, and Bob Mankoff. A kl-lucb bandit algorithm for large-scale crowdsourcing. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5896–5905, 2017.

Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration for interactive machine learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019,*

December 8-14, 2019, Vancouver, BC, Canada, pages 2887–2897, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4f398cb9d6bc79ae567298335b51ba8a-Abstract.html.

Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.

Yanhong Wu. *Inference for change point and post change means after a CUSUM test*, volume 180. Springer Science & Business Media, 2007.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262. PMLR, 2016.