ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230465

Piecewise-Stationary Combinatorial Semi-Bandit with Causally Related Rewards

Behzad Nourani-Koliji, Steven Bilaj, Amir Rezaei Balef and Setareh Maghsudi

University of Tübingen
ORCiD ID: Behzad Nourani-Koliji https://orcid.org/0009-0001-2854-1955

Abstract. We study the piecewise stationary combinatorial semi-bandit problem with causally related rewards. In our nonstationary environment, variations in the base arms' distributions, causal relationships between rewards, or both, change the reward generation process. In such an environment, an optimal decision-maker must follow both sources of change and adapt accordingly. The problem becomes aggravated in the combinatorial semi-bandit setting, where the decision-maker only observes the outcome of the selected bundle of arms. The core of our proposed policy is the Upper Confidence Bound (UCB) algorithm. We assume the agent relies on an adaptive approach to overcome the challenge. More specifically, it employs a change-point detector based on the Generalized Likelihood Ratio test. Besides, we introduce the notion of *group restart* as a new alternative restarting strategy in the decision making process in structured environments. Finally, our algorithm integrates a mechanism to trace the variations of the underlying graph structure, which captures the causal relationships between the rewards in the bandit setting. Theoretically, we establish a regret upper bound that reflects the effects of the number of structural- and distribution changes on the performance. The outcome of our numerical experiments in real-world scenarios exhibits applicability and superior performance of our proposal compared to the state-of-the-art benchmarks.

1 Introduction

Multi-armed bandit (MAB) [27] is a class of sequential learning- and optimization problems. In the seminal MAB problem, the decision-maker (agent) selects one of the K available arms, where each arm returns a reward drawn from a time-invariant, unknown distribution. The agent maximizes the total expected reward over the gambling horizon by using an effective decision-making strategy that maps the historical actions and outcomes to future actions. That is equivalent to minimizing the total expected *regret*, which is the difference between the reward of the applied policy and that of the optimal policy in hindsight. Indeed, the MAB challenge boils down to the exploration-exploitation dilemma, where the agent decides between accumulating immediate rewards on the one side and obtaining information that might result in a larger reward only in the future on the other side. Due to its wide variety, the MAB framework is a potential candidate as a mathematical tool for tackling many real-world problems, for example, resource allocation in networks [23], recommender systems [20], and clinical trials [3].

The combinatorial multi-armed bandit (CMAB) problem is an extension of the seminal MAB. Instead of only one arm in each round, the agent chooses a number of them, i.e., it takes a combinatorial action. That results in exponential growth of the decision set by increasing the number of arms. Consequently, the conventional MAB methods such as UCB1 [1] become inefficient or inapplicable. In CMAB, we refer to each original arm as a base arm, and any subset of the base arms is a *super arm*. Sometimes, the agent observes the reward of all base arms inside the super arm; In some other cases, the agent observes only one reward. The former type of feedback is a *semi-bandit feedback*, whereas the latter is a *bandit feedback*. The bandit problem becomes aggravated when a statistical structure influences the reward generation processes so that besides the excessively-large action set, the player deals with the structural relationships to decide optimally. We focus on combinatorial semi-bandit (CSB) problem with causally related rewards.

The seminal settings of CMAB- or CSB problems do not assume any statistical or probabilistic relationship between the base arms; Nevertheless, in several application domains, the potential dependency between the random variables can be abstracted by a structure. Despite being neglected for a long time, different types of the MAB problem with probabilistic or statistical relationships between the base arms, referred to as *structured bandits* receive increasing attention from the research community in the past few years. For example, [10], [35], and [18] assume that some arms may probabilistically be triggered based on the outcome of other arms. In [19], prior knowledge about the causal structure that affects the rewards is available. The authors in [26] introduce a causally structured CSB problem and use a directed acyclic graph to model the causal structure that influences the reward generation process. Their algorithm does not need apriori knowledge concerning the structural relationships as it can learn the structure from the streaming data. All of the works mentioned above study a stationary setting.

Unlike the stationary stochastic setting, in many real-world scenarios, the reward distributions of base arms change over time in an evolving environment. For example, in recommender systems, the behavioral feedback of users is time-variant. It is possible to address the nonstationary behaviors of rapidly-varying environments using the adversarial bandit framework [2]. However, in some cases, the environment changes slowly and less frequently. In such scenarios, policies designed for stationary or adversarial bandits are sub-optimal. Generally speaking, there are two main approaches in modeling this type of nonstationarity in bandit problems; the *switching case* (abruptly changing) [37] and the *dynamic case* (smoothly changing) [9] [32]. For the switching case, the reward distributions of base arms remain unchanged for certain intervals. The en-

Please check ArXiv or contact the authors for any appendices or supplementary material mentioned in the paper.

vironment then varies if the distributions of a subset of base arms change instantly. The point where distributions change is a *change-point* (or breakpoint) and an interval between any two consecutive change-point is a *stationary segment* [37]. In contrast, in the dynamic case, the base arms' mean rewards evolve slowly instead of abruptly changing at one point, and the variation is bounded by a variation budget [5]. In this paper, we focus on the switching case, also referred to as *piecewise* stationary bandit model [6]. We measure the decision-making performance using the notion of *piecewise stationary regret*, i.e., the regret w.r.t. an oracle that knows the best action in each stationary segment.

In a piecewise stationary structured bandit problem, the reward generation processes might vary by changing the base arms' reward distributions and the structural relationships between the variables. Although such a model has remained unaddressed in the MAB literature, it accommodates several real-world applications. Those include financial markets, where not only the *investors' stock purchasing behavior* but also the *causal effects amongst the stock prices* can be time-varying [30]. In such scenarios, an optimal investor follows both sources of change and adapts accordingly. While the availability of prior knowledge about the structural relationships is a strong and unrealistic assumption, inferring such structural relationships from the streaming partial feedback in the bandit setting is also challenging. We study a piecewise stationary structured CSB problem, where the causal relationships between the rewards and the distributions of the base arms evolve. In order to model the structural relationships, we rely on *Structural Equation Models* [15].

In general, there are two main approaches to follow the piecewise stationary behaviour of base arms distributions; *passively adaptive* approach [13] and *actively adaptive* approach [6] [8] [16] [22] [37] [11]. Methods of the former category are unaware of the change-points and rely on their understanding of the optimal action based on the most recent observations. On the contrary, methods of the latter category use a change detection algorithm to follow the distributions' changes and decide accordingly [6]. Some studies show the superior performance of actively adaptive approaches [24]. Clearly, the performance of actively adaptive approaches rely significantly on the ability of the agent in handling the breakpoints. Current actively adaptive algorithms incorporate either *global restart* or *local restart* to restart learning the expected value of the instantaneous rewards of the base arms. The former method resets learning the expected values of all arms after detecting a change in one of them. The latter restarts learning only for those arms undergoing a change. These approaches suffer from a drawback as they ignore possible relationships amongst arms' distributions in making a decision upon restarting process. There are main reasons for introducing a new restarting strategy for bandit algorithms in structured piecewise stationary environments. Firstly, social networks, as one of the main target applications for bandit algorithms, exhibit large modularity measures [25] [7]. Secondly, in some real-world scenarios changes within a network are not completely independent, but they are rather the result of the local spread of a change-seed within the network structure through mechanisms such as contagion [12], social influence [12], or diffusion [31], e.g. media-based marketing campaigns, or rumor diffusion over social networks [29]. In this regard, we introduce the notion of *group restart* where we restart the set of arms that are in the same group, upon detecting a change in any of them. We elaborate more on this in the following sections. We show the superior adaptation capabilities of this approach over local and global restarts in our experiments and discuss the effects of this approach over the upper bound of regret in theory.

In this work, we introduce a piecewise stationary CSB problem with causally related rewards. Our framework accommodates the changes in the base arms' reward distributions and also in the causal relationships between the rewards. We provide an actively adaptive approach to tackle the problem. We introduce a novel alternative restarting strategy, namely *group restart*, that can be used in the adaptation of stationary bandit algorithms to the piecewise stationary environments. We highlight the importance of using the knowledge of relationships amongst arms' distributions in our group restart strategy. We achieve this by showing its effects on the regret of the algorithm in dealing with the costly effects of both *restarts* and *delays of change point detectors*. Our algorithm uses a UCB-based policy for learning the expected rewards of the base arms and a Generalized Likelihood Ratio (GLR) change-point detector. Furthermore, we integrate a mechanism in our algorithm to follow the changes of the causal graph structure that models the causal relationships between the rewards in the bandit setting. We provide the theoretical analysis of the regret upper bound for our algorithm. Our regret bound reflects the effects of both the number of causal graph changes and the number of distribution changes. Our numerical experiments using synthetic- and real-world data establish the advantage of our algorithm compared to the benchmarks.

In **Section 2**, we introduce the piecewise stationary combinatorial semi-bandit problem with causally related rewards. In **Section 3**, we develop our decision-making policy, namely, PS-SEM-UCB-Gr. **Section 4** presents the theoretical analysis of the regret performance of PS-SEM-UCB-Gr. **Section 5** includes the numerical experiments. **Section 6** concludes the paper with some suggestions for future works.

2 Problem Formulation

In a **piece-wise stationary combinatorial semi-bandit (PSCSB)** problem with causally related rewards, a change from one stationary segment to the other results from varying (i) base arms' reward distributions or (ii) the causal relationships between rewards. The intervals with fixed reward distributions and static causal graph are distribution- and graph stationary segments, respectively. The change-points of both segment types appear randomly. We use $\mathcal{K} = \{1, \ldots, \mathrm{K}\}$ to represent the set of K base arms, $\mathcal{D} \subseteq 2^{\mathcal{K}}$ the set of all super arms, and $\mathcal{T} = \{1, \ldots, T\}$ a sequence of T time-steps. Besides, $\theta_{k,t}$ is the distribution of the instantaneous reward of arm k at time t with mean $\mu_{k,t}$ and bounded support within $[0, 1]$. The vector $\boldsymbol{\mu}_t = [\mu_{1,t}, \ldots, \mu_{K,t}]$ is the expected values of the instantaneous rewards of all base arms at time t. Additionally, $\mathcal{A}_t$ is the underlying graph that shows the causal relations between the base arms' rewards. We use $\psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\mathbf{x}_t)$ to denote the agent's expected reward from the decision vector $\mathbf{x}_t$ given $\boldsymbol{\mu}_t$ and $\mathcal{A}_t$. Consequently, we characterize a PSCSB with the tuple $(\mathcal{K}, \mathcal{D}, \mathcal{T}, \{\theta_{k,t}\}_{k \in \mathcal{K}, t \in \mathcal{T}}, \psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\mathbf{x}_t))$. Vector of base arms' instantaneous rewards at time t is represented by $\mathbf{b}_t = [b_{1,t}, \ldots, b_{K,t}] \in [0, 1]^K$ and it follows a piece-wise independent and identically distributed (i.i.d.) model in each distribution stationary segment. A change to the distribution stationary segment of the environment corresponds to a change in at least one arm's reward distribution. Our setting assumes that the agent is given the meta information regarding the grouping (clustering) of arms such that arms within the same group tend to have their instantaneous rewards' distributions changed together. We use g to denote a group of arms. $K_g = |g|$ is used to show the cardinality of the group g. g_k represents the group to which arm k belongs. We use G to denote the set of all groups, $G = \left\{g^{(1)}, \ldots, g^{(\zeta)}\right\}$, with

$|G| = \zeta$ and $\bigcup_{i\in[\zeta]} g^{(i)} = \mathcal{K}, \forall i, j \in [\zeta], g^{(i)} \cap g^{(j)} = \varnothing$ where $[\zeta] = \{1, \ldots, \zeta\}$. N_Θ is used to denote the number of distribution stationary segments of the environment. We define the total number of distribution stationary segments for group g as

$$N_g = 1 + \sum_{t=1}^{T-1} \mathbb{1}\left\{\exists k \in g \text{ s.t. } \theta_{k,t} \neq \theta_{k,t+1}\right\}. \tag{1}$$

Hence, the total number of stationary segments for all groups is $N_G = \sum_{g\in G} N_g$. This clarifies that N_G can change depending on the way the grouping is performed. At each time t, the agent selects a *decision vector* $\mathbf{x}_t = [x_{1,t}, \ldots, x_{K,t}] \in \{0,1\}^K$. We use $\mathcal{I}_t \subset \mathcal{K}$ to denote the set of chosen base arms $I_t \in \mathcal{K}$ at round t. We have $x_{k,t} = 1$ if the base arm k is in the super arm $\mathcal{I}_t$ at time t, otherwise $x_{k,t} = 0$. The agent selects at most m base arms at each time step. Hence, we define the set of all feasible decision vectors as $\mathcal{X} = \left\{\mathbf{x} \mid \mathbf{x} \in \{0,1\}^K \wedge \|\mathbf{x}\|_0 \leq m\right\}$ where $\|\cdot\|_0$ determines the number of non-zero elements in a vector and the parameter m is pre-determined. The causal relationships in the environment are modelled using a directed graph. More precisely, we consider an unknown piecewise static sparse Directed Acyclic Graph (DAG), $\mathcal{A}_t = (\mathcal{V}, \mathcal{E}_t, \mathbf{W}_t)$. $\mathcal{V}$ represents the set of K vertices, i.e., $|\mathcal{V}| = K$, $\mathcal{E}_t$ and $\mathbf{W}_t$ denote the edge set and the weighted adjacency matrix at time t, respectively. We allow the edge set $\mathcal{E}_t$ to change arbitrarily every time the causal graph structure changes. However, the set of vertices $\mathcal{V}$ stays unchanged across time. This implies that the adjacency matrix $\mathbf{W}_t$ changes only in the elements $\mathbf{W}_t[i,j], \forall i, j \in \mathcal{K}, \forall t \in \mathcal{T}$, as far as the underlying graph structure remains a DAG without self-loop, i.e., $\mathbf{W}_t[i,i] = 0, \forall i \in \mathcal{K}, \forall t \in \mathcal{T}$. $N_{\mathbf{W}}$ represents the number of graph stationary segments. An error-free piecewise static Structural Equation Model (SEM) [15] is used to model the generation of reward in the environment. At each time t, $\mathbf{z}_t = [z_{1,t}, \ldots, z_{K,t}]$ is used to represent the exogenous input vector while $\mathbf{y}_t = [y_{1,t}, \ldots, y_{K,t}]$ denotes the endogenous output vector of the SEM. We write,

$$\mathbf{z}_t = \mathrm{diag}(\mathbf{b}_t)\mathbf{x}_t, \tag{2}$$

where $\mathrm{diag}(\cdot)$ represents a diagonal matrix. This implies that the exogenous input $\mathbf{z}_t$ contains the semi-bandit feedback in the decision-making problem. We define the k^{th} element of the endogenous output vector $\mathbf{y}_t$ at any time t as

$$y_{k,t} = \sum_{j=1}^{K} \mathbf{W}_t[k,j] y_{j,t} + z_{k,t}, \quad \forall k \in \mathcal{K}, \tag{3}$$

At each time t, the endogenous output $y_{k,t}$ represents the *overall reward* of base arm $k \in \mathcal{K}$. The element $\mathbf{W}[k,j]$ represents the causal effect of the overall reward of base arm j on the overall reward of base arm k. Therefore, the overall rewards of base arms are causally related while the instantaneous reward of arm k only directly contributes to the overall reward of arm k. It is important to distinguish between the relationships amongst arms' distributions and the causal relationships amongst the overall rewards. The first one only explains the prior information regarding the groupings of arms, while the second one is used in the mathematical formulation of the problem.

The adjacency matrices $\mathbf{W}_t, \forall t \in \mathcal{T}$ are unknown a priori and $\mathbf{W}_t[i,j] \geq 0, \forall i, j \in \mathcal{K}, \forall t \in \mathcal{T}$. The matrix form of (3) at time t is given as

$$\mathbf{y}_t = \mathbf{W}_t \mathbf{y}_t + \mathbf{z}_t. \tag{4}$$

As a result, we write $\mathbf{y}_t = (\mathbf{I} - \mathbf{W}_t)^{-1} \mathrm{diag}(\mathbf{b}_t)\mathbf{x}_t$ by solving (4) for $\mathbf{y}_t$, where $\mathbf{I}$ is the identity matrix. We assume that the agent is able to observe both the instantaneous semi-bandit feedback vector $\mathbf{z}_t$ and the overall reward feedback vector $\mathbf{y}_t$. The *payoff* received by the agent upon choosing the decision vector $\mathbf{x}_t$ is defined as

$$r_t(\mathbf{x}_t) = \mathbf{c}^\top \mathbf{y}_t = \mathbf{c}^\top (\mathbf{I} - \mathbf{W}_t)^{-1} \mathrm{diag}(\mathbf{b}_t)\mathbf{x}_t, \tag{5}$$

where $\mathbf{c} = [c_1, \ldots, c_K] \in \{0,1\}^K$ is pre-determined. The agent is interested in the output y_k in the causal network if $c_k = 1$, and $c_k = 0$ otherwise. Since the graph $\mathcal{A}_t$ is a DAG, the adjacency matrix $\mathbf{W}_t$ is nilpotent. This property guarantees that the matrix $(\mathbf{I} - \mathbf{W}_t)$ is invertible. Given a decision vector $\mathbf{x}_t \in \mathcal{X}$, the expected payoff at time t is calculated as

$$\psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\mathbf{x}_t) = \mathbb{E}\left[r_t(\mathbf{X}) | \mathbf{X} = \mathbf{x}_t\right], \tag{6}$$

where the expectation concerns the randomness in the reward generating process. We denote by $\mathbf{x}_t^* = \underset{x\in X}{\mathrm{argmax}}\ \psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\mathbf{x})$ the decision vector with maximum expected reward at time t. The agent minimizes the cumulative piecewise stationary regret defined as

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{t=1}^{T} (\psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\mathbf{x}_t^*) - \psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\mathbf{x}_t))\right]. \tag{7}$$

3 The Learning Algorithm

In this section, we develop a solution to the formulated problem. We first introduce the group restart strategy, and the online graph learning. Afterward, we present our decision-making policy, namely, PS-SEM-UCB-Gr.

3.1 Group Restart Strategy

Restarting process plays a key role in the decision making strategy in piecewise stationary bandit algorithms. Upon taking the global restart strategy, the agent's regret increases due to the costly effects of restarting of all arms. Moreover, by taking local restart strategy, delays of change point detectors for different arms can make the algorithm to incur linear regret in some intervals. One way to address these issues is in the structured environments where changes are not always completely independent and having side information w.r.t. relationships between arms' distributions can be helpful in making decisions upon restarts. There are certain research directions in MAB literature where relationships amongst the arms are considered. For instance, in [33], it is assumed that each item that the algorithm recommends is a node of a known graph and the expected rating of the neighboring nodes are similar. Furthermore, in [14], it is suggested that the nodes of the graph can be clustered according to some apriori unknown clustering and the arms within the same cluster exhibit similar behaviours. Also, in [36], the relationship between the users is captured by an underlying graph and user preferences are assumed to have smooth signals on the graph. In such settings, it is natural to anticipate that if an arm's expected reward is changed, then due to the relationships of the arms, the set of arms that are closely connected to it go through changes as well. Consequently, we propose *group restart* strategy as an efficient alternative in structured environments where grouping information might be either available in advance or learned from the data [14] [21]. As the result of our theoretical analysis, we show that a structure-based grouping in group restart strategy can help to reduce the regret upper bound compared to local and global restarts.

Algorithm 1 Graph Learning Data Generation (GLDG)

Create an initialization matrix **init**, $\mathbf{Y}_0 = []$, $\mathbf{Z}_0 = []$.
for $t' = 1, 2, \ldots, K$ **do**
  $\mathbf{x}_t := \mathbf{init}[:, t']$
  Play $\mathcal{I}_t$, receive reward $r(\mathbf{x}_t)$, $s_{I_t,n_{I_t,t}} \leftarrow z_{I_t,t}, \forall I_t \in \mathcal{I}_t$.
  for all $I_t \in \mathcal{I}_t$ **do**
    update: $\hat{\mu}_{I_t,t}$ using (9), $n_{I_t,t}$ using (10).
    if GLR$(s_{I_t,1}, \ldots, s_{I_t,n_{I_t,t}}; \delta) = 1$ **then**
      $\forall k \in g_{I_t}$: $n_{k,t} \leftarrow 0$, $\hat{\mu}_{k,t} \leftarrow 0$, $\tau_k \leftarrow t$.
      $\tau' \leftarrow t$, $\Omega \leftarrow \Omega \cup g_{I_t}$.
  for all $k \in \mathcal{K}$ **do**
    if $n_{k,t} \neq 0$ **then**
      $U_{k,t} \leftarrow \hat{\mu}_{k,t} + \sqrt{\frac{(m+1)\log(t-\tau_k)}{n_{k,t}}}$
  $[\mathbf{Y}_t] \leftarrow [\mathbf{Y}_{t-1}, \mathbf{y}_t]$, $[\mathbf{Z}_t] \leftarrow [\mathbf{Z}_{t-1}, \mathbf{z}_t]$, $t \leftarrow t+1$
Solve (8) to get $\hat{\mathbf{W}}_{t-1}$.
flag $= 0$

3.2 Piece-wise Static Graph Learning

Considering the required knowledge of $\mathbf{W}_t$ in finding the optimal decision vector, we propose an online graph learning framework that uses the collected feedback $\mathbf{y}_t$ and $\mathbf{z}_t$ and allows for modelling both the random and the smooth transitions of the causal graph. At each time t, we stack the feedback, from the last graph-change-point up to the current time, as consecutive columns in $\mathbf{Z}_t$ and $\mathbf{Y}_t$, hence, $\mathbf{Y}_t = \mathbf{W}_t\mathbf{Y}_t + \mathbf{Z}_t$. We use the collected feedback history, $\mathbf{Y}_t$ and $\mathbf{Z}_t$, as the input to a parametric graph learning algorithm for a static SEM [15]. Formally, the adjacency matrix at time t is the solution to the following optimization problem:

$$\begin{aligned} \hat{\mathbf{W}}_t = \underset{W \in R^{K\times K}}{\operatorname{argmin}} \quad & \|\mathbf{Y}_t - \mathbf{W}\mathbf{Y}_t - \mathbf{Z}_t\|_{\mathrm{F}}^2 + \lambda_1 \|\mathbf{W}\|_1 \\ \text{s.t.} \quad & \mathbf{W}[k,k] = 0, \forall k \in \mathcal{K} \end{aligned} \tag{8}$$

where $\|\cdot\|_{\mathrm{F}}$ represents the Frobenius norm of matrices. The symbol $\|\cdot\|_1$ denotes the L^1-norm of the matrices and it is used to impose sparsity over the estimated adjacency matrix $\hat{\mathbf{W}}_t$. We use the notation $\hat{\mathbf{W}}^{(i)}$ to represent the estimated adjacency matrix for the i^{th} static graph. In order to impose slow topological variations across time, from one static graph to the next, one may add a second regularization term $\lambda_2 \left\|\hat{\mathbf{W}}^{(i+1)} - \hat{\mathbf{W}}^{(i)}\right\|_1$ in (8) and have a form of the optimization problem in (8) that stays convex. This second regularization allows the algorithm to penalize deviation of the current graph estimate from the predecessor, hence implementing a transfer of knowledge that is gained from the previous segment.

3.3 The PS-SEM-UCB-Gr Algorithm

In this section, we describe our decision-making policy. Its core is the Upper Confidence Bound policy. Besides, we use two previously-proposed methods, namely *group restart*, and *piece-wise static graph learning*. Finally, we integrate a mechanism for detecting the changes to the adjacency matrix of the causal graph. Each time the algorithm decides to infer the new adjacency matrix, it starts a subroutine inside the main algorithm to obtain K data samples by interacting with the new environment. It is crucial that the new dataset satisfies the conditions for the precise inference and unique identification of the new graph adjacency matrix [4] [26]. We refer to this subroutine as *Graph Learning Data Generation* (GLDG). For these K rounds, PS-SEM-UCB-Gr picks K columns of an *initialization matrix*, namely,

Algorithm 2 PS-SEM-UCB-Gr: Piecewise Stationary - Structural Equation Model - Upper Confidence Bound - Group Restart

1: **Initialization:** $\forall k \in \mathcal{K}$, $n_{k,0} \leftarrow 0$, $\hat{\mu}_{k,0} \leftarrow 0$, $\tau_k = 0$, $t = 1$, $\tau' = 0$, *flag* $= 1$. Get $G = \left\{g^{(1)}, \ldots, g^{(\zeta)}\right\}$
2: **while** $t \leq T$ **do**
3: **if** *flag* $= 1$ **then**
4: Run ***GLDG***.
5: **if** $\Omega \neq \varnothing$ **then**
6: Pick $a \in \Omega$, Randomly choose $\mathcal{I}_t$ with $a \in \mathcal{I}_t$.
7: Remove a from Ω.
8: **else**
9: Solve (11) for $\mathbf{x}_t$.
10: Play $\mathcal{I}_t$, receive reward $r(\mathbf{x}_t)$, $s_{I_t,n_{I_t,t}} \leftarrow z_{I_t,t}, \forall I_t \in \mathcal{I}_t$.
11: **for all** $I_t \in \mathcal{I}_t$ **do**
12: update: $\hat{\mu}_{I_t,t}$ using (9), $n_{I_t,t}$ using (10).
13: **if GLR**$(s_{I_t,1}, \ldots, s_{I_t,n_{I_t,t}}; \delta) = 1$ **then**
14: $\forall k \in g_{I_t}$: $n_{k,t} \leftarrow 0$, $\hat{\mu}_{k,t} \leftarrow 0$, $\tau_k \leftarrow t$.
15: $\tau' \leftarrow t$, $\Omega \leftarrow \Omega \cup g_{I_t}$.
16: **if** $\exists\, c \in \mathbb{N} : t - \tau' = c \left\lfloor \frac{K}{p} \right\rfloor$ **then**
17: $\Omega = \bigcup_{i\in[\zeta]} g^{(i)}$
18: **for all** $k \in \mathcal{K}$ **do**
19: **if** $n_{k,t} \neq 0$ **then**
20: $U_{k,t} \leftarrow \hat{\mu}_{k,t} + \sqrt{\frac{(m+1)\log(t-\tau_k)}{n_{k,t}}}$
21: $[\mathbf{Y}_t] \leftarrow [\mathbf{Y}_{t-1}, \mathbf{y}_t]$, $[\mathbf{Z}_t] \leftarrow [\mathbf{Z}_{t-1}, \mathbf{z}_t]$
22: **if** $\left\|\mathbf{y}_t - \hat{\mathbf{W}}_{t-1}\mathbf{y}_t - \mathbf{z}_t\right\|_2^2 > \epsilon$ **then**
23: *flag* $= 1$, $\mathbf{Y}_t = []$, $\mathbf{Z}_t = []$
24: **else**
25: Solve (8) to get $\hat{\mathbf{W}}_t$.
26: $t \leftarrow t + 1$

$\mathbf{Init} \in \{0,1\}^{K\times K}$ in a sequential way where $\mathbf{Init}$ is created as described in [26], Section 3.2. Based on the discussion above, we assume that there are at least $K+1$ rounds between any two consecutive changes in the graph. That guarantees sufficient time to infer the new ground truth graph after every change. We refer to the rounds inside a GLDG phase as *graph initialization* rounds and the rest as *normal* rounds. In every round, the GLR change-point detectors and the UCB index developments are working. The input parameters of PS-SEM-UCB-Gr include the number of steps (T), number of arms (K), uniform exploration probability $p \in (0,1)$, and δ as the confidence level of the GLR change-point detector. The policy uses the parameter τ' to perform the uniform forced exploration over all base arms in Line 16 of Algorithm 2. The forced uniform exploration guarantees that the GLR change-point detectors receive sufficient samples. Considering that we are using group restart, UCB developments of arms from different groups might have different resetting times. Therefore, the policy uses $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_K]$ to manage the restarting times of UCB indices. The variable *flag* is used to call the GLDG subroutine. For any arm k, the empirical average of the instantaneous rewards at any time $t = t_1$ w.r.t. its last restarting time at $t = \tau_k$ yields

$$\hat{\mu}_{k,t_1} = \frac{\sum_{t=\tau_k+1}^{t_1} z_{k,t}}{n_{k,t_1}}, \tag{9}$$

where n_{k,t_1} is the number of times that the base arm k is observed up to time $t = t_1$ since its last restart at $t = \tau_k$. Formally,

$$n_{k,t_1} = \sum_{t=\tau_k+1}^{t_1} x_{k,t}. \tag{10}$$

The set Ω holds the index of those arms whose UCB developments are being restarted or that are candidates for forced exploration. After the graph initialization period, in each round, PS-SEM-UCB-Gr first checks the set Ω, in Line 5, otherwise the agent plays the next super arm according to the result of the combinatorial optimization in Line 9. The combinatorial optimization uses the current UCB indices and the last estimate of the causal graph. We denote the UCB index of base arm k at time t as $U_{k,t}$ such that we have the UCB indices of all base arms in the vector $\mathbf{U}_t = [U_{1,t}, \ldots, U_{K,t}]$. Therefore, the combinatorial optimization for finding the best decision vector yields

$$\mathbf{x}_t = \underset{x \in X}{\operatorname{argmax}} \quad \mathbf{c}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \operatorname{diag}(\mathbf{U}_{t-1})\mathbf{x}. \tag{11}$$

Let $\mathbf{M}^\top = \mathbf{c}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \operatorname{diag}(\mathbf{U}_{t-1})$. The elements of $\hat{\mathbf{W}}_{t-1}$, $\mathbf{c}$, and $\mathbf{U}_{t-1}$ are non-negative, then the optimization problem (11) can be solved by finding a subset of elements in $\mathbf{M}$ such that $\mathbf{x} \in \mathcal{X}$. Therefore, it is solvable by using an efficient sorting algorithm that ranks the elements of $\mathbf{M}$. Consequently, the agent plays $\mathbf{x}_t$, collects the reward in Line 10 according to (5), and updates the vectors $\hat{\boldsymbol{\mu}}_t$ and $\mathbf{n}_t$ in Line 12. The notation $s_{I_t,n_{I_t,t}} \leftarrow z_{I_t,t}$ in Line 10 implies that the collected feedback $z_{I_t,t}, \forall I_t \in \mathcal{I}_t$ is the sample number $n_{I_t,t}$ in the sequence of samples for arm I_t since its last restart at $t = \tau_{I_t}$. We use the **GLR** change-point detector [6] defined as

$$\begin{aligned}\mathbf{GLR}(s_1, \ldots, s_n; \delta) \coloneqq \\ \mathbb{1}\{\sup_{\alpha \in [1,n-1]} [\alpha \times \mathrm{kl}(\hat{\beta}_{1:\alpha}, \hat{\beta}_{1:n}) + \\ (n-\alpha) \times \mathrm{kl}(\hat{\beta}_{\alpha+1:n}, \hat{\beta}_{1:n})] \geq \gamma(n,\delta)\},\end{aligned} \tag{12}$$

where $\hat{\beta}_{\alpha:\alpha'}$ is the mean of the observations between α and α', $\mathrm{kl}(x,y) = x \log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right)$ is the binary relative entropy between any two Bernoulli distributions with means x and y. The function $\gamma(n,\delta)$ is the threshold function for the GLR test. Theoretically, we choose this threshold function following Lemma 2 in [6]. However, in all our numerical experiments, we follow "Practical considerations" in [6], and select $\gamma(n,\delta) = \ln(\frac{3n\sqrt{n}}{\delta})$. If $\mathbf{GLR}(s_1, \ldots, s_n; \delta) = 1$, the algorithm applies group restarts in Line 14. The algorithm updates the UCB indices in Line 20. In Line 22, the graph-change detection mechanism uses two vectors of $\mathbf{z}_t$ and $\mathbf{y}_t$ to test the validity of the last estimate of the graph adjacency matrix, $\hat{\mathbf{W}}_{t-1}$. If the error value for the graph-change detection formulation exceeds ϵ, then the algorithm notifies that the previous estimate of the graph structure is no longer valid. Consequently, in Line 23, we have flag $= 1$, and the previously collected sets of feedback in $\mathbf{Z}_t$ and $\mathbf{Y}_t$ are dropped. Parameter ϵ represents the error we accept in the graph estimation process. In this paper, we take $\epsilon = 0$. However, assuming $\epsilon \neq 0$, the effects of ϵ should be considered in the regret analysis. In case the collected feedback vectors $\mathbf{y}_t$ and $\mathbf{z}_t$ satisfy the SEM formulation for $\hat{\mathbf{W}}_{t-1}$, in Line 21, PS-SEM-UCB-Gr uses the newly updated matrices $\mathbf{Y}_t$ and $\mathbf{Z}_t$, to improve the adjacency matrix estimation. It is important to notice that the algorithm does not restart the UCB development upon detecting a graph-change. It also does not restart the graph learning following any distribution-change detection.

4 Theoretical Analysis

In this section, we deliver the analysis for the expected regret of PS-SEM-UCB-Gr algorithm. We perform the regret analysis according to any grouping of arms, with local and global restarts as special cases. We denote the maximum delay across all detected changes as d. We divide the time line into stationary segments of base arm distributions. The graph changes will be treated separately as they do not affect the UCB developments and only contribute to the regret in terms of a constant, based on the graph-learning phase. We also define the suboptimality gaps in our setting as the reward difference between the optimal decision vector $\mathbf{x}^*$ and an arbitrary decision vector $\mathbf{x}$: $\Delta_t(\mathbf{x}) = \psi_t(\mathbf{x}_t^*) - \psi_t(\mathbf{x})$, where $\psi_t(\mathbf{x})$ is the mean reward of $\mathbf{x}$, with only subscript t used to parameterize it for better readability. The largest gap is denoted as $\Delta_{\max} = \max_t \max_{\mathbf{x}:\psi_t(\mathbf{x})<\psi_t(\mathbf{x}_t^*)} \Delta_t(\mathbf{x})$, and the smallest $\Delta_{\min} = \min_t \min_{\mathbf{x}:\psi_t(\mathbf{x})<\psi_t(\mathbf{x}_t^*)} \Delta_t(\mathbf{x})$. As it is essential for estimating the total regret bound, we deliver the regret for the stationary case, with an improvement over the work of [26], as our result does not scale with the number of layers in the causal graph but only with the total number of arms;

Lemma 1 *Let* $\omega_t^T = \mathbf{c}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \operatorname{diag}(\mathbf{x}_{t+1})$ *and* $\omega_{\max} = \max_t \max_k \omega_{k,t}, k \in \mathcal{K}$. *In the stationary case* ($N_\Theta = 1 \wedge N_{\mathbf{W}} = 1$) *of the PS-SEM-UCB-Gr algorithm, the upper regret bound is given as:*

$$\mathcal{R}(T) \leq \left[\frac{4\omega_{\max}^2 m^2 (m+1) K \log(T)}{\Delta_{\min}^2} + \frac{\pi^2}{3} mK + K\right] \Delta_{\max},$$

with $\Delta_{\max}$ *as the largest suboptimality gap and* $\Delta_{\min}$ *smallest suboptimality gap.*

The adapted proof is given in the supplementary materials. The following Theorem 1 states a bound on the regret in the non-stationary case of our proposed decision-making policy for any grouping of arms.

Theorem 1 *Let* $\omega_t^T = \mathbf{c}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \operatorname{diag}(\mathbf{x}_{t+1})$ *and* $\omega_{\max} = \max_t \max_k \omega_{k,t}, k \in \mathcal{K}$. *The expected regret of the PS-SEM-UCB-Gr policy is upper bounded as:*

$$\begin{aligned}\mathcal{R}(T) \leq \sum_{g \in G} \left[N_g K_g R_0(T) + (\delta T + 1 + \frac{\pi^2 m}{3}) N_g K_g \Delta_{\max}\right] \\ + (Tp + dN_G + \delta T(K + N_G) + N_{\mathbf{W}} K)\, \Delta_{\max},\end{aligned}$$

with $R_0(T) = \frac{4\omega_{\max}^2 m^2 (m+1) \log(T)}{\Delta_{\min}^2} \Delta_{\max}$.

See Section 1 of supplementary material for the proof.

This is the general regret upper bound that reflects the importance of grouping of arms. We are able to retrieve the bounds according to the given grouping of arms. We assumed the knowledge of groupings of base arms based on structural relationships between arms' distributions.

Following local restart strategy, $G = G_{\text{local}}$, we have $K_g = 1$, $\forall g \in G_{\text{local}}$ and $|G_{\text{local}}| = K$, thus $\sum_g N_g K_g = N_{G_{\text{local}}}$. If we follow global restart strategy, $G = G_{\text{global}}$, we have $K_g = K, \forall g \in G_{\text{global}}$ and $|G_{\text{global}}| = 1$, thus $\sum_g N_g K_g = K N_{G_{\text{global}}}$. It is important to note that the number of restarts differs for local and global restart strategies, since $N_{G_{\text{global}}} \leq N_{G_{\text{local}}}$. In the following, we compare the performance of our approach with local restarts and global restarts on the amount of regret increase in the distribution stationary segment after a breakpoint.

Remark 1 *We rewrite the regret upper bound in Theorem 1 as* $R(T) \leq \Sigma_{g\in G}[C_1 N_g K_g + C_2 N_g] + C_3$ *where* C_1, C_2, C_3 *are independent of the grouping of arms. Let us assume that the breakpoint* ν *happens from* t *to* $t+1$ *with change to* $\mathfrak{K}_\nu$ *arm distributions that belong to* η_ν *groups (clusters). The increase of the regret value within the stationary segment after breakpoint* ν *can be written as* $\Delta R(\nu) \leq C_1 \Sigma_{g\in G} K_g \mathbb{1}\{\exists k \in g \text{ s.t. } \theta_{k,t} \neq \theta_{k,t+1}\} + C_2 \Sigma_{g\in G} \mathbb{1}\{\exists k \in g \text{ s.t. } \theta_{k,t} \neq \theta_{k,t+1}\}$. *Consequently, we have the followings;*

- *In the case of Local restart, we have* $\Delta R(\nu) \leq C_1 \mathfrak{K}_\nu + C_2 \mathfrak{K}_\nu$.
- *In the case of Global restart, we have* $\Delta R(\nu) \leq C_1 K + C_2$.
- *If the total number of arms inside the* η_ν *groups is* $\mathfrak{K}_\nu$*, for Group restart, we have* $\Delta R(\nu) \leq C_1 \mathfrak{K}_\nu + C_2 \eta_\nu$.
- *If in the* η_ν *groups, there are collectively* s *arms whose distributions did not change at* ν*, in this case, for Group restart we have* $\Delta R(\nu) \leq C_1(\mathfrak{K}_\nu + s) + C_2 \eta_\nu$.

In the above, the first term, scaling with C_1, is the regret due to number of restarted arms, while the second term, scaling with C_2, is affected by the delays. These results clarify the idea behind using a group restart strategy, especially in cases where the $\mathfrak{K}_\nu$ changed arms are from a small number of η_ν clusters. Intuitively, in networks with high modularity measures, we can expect to have smaller number for s and a better performance for the group restart strategy.

By the following corollary, through fine-tuning the hyper-parameters δ and p and with the assumption of the prior knowledge of N_G, we can achieve a sub-linear regret bound;

Corollary 1 *Let* $\Delta_{\min}^{change} = \min_i \max_{k\in\mathcal{K}} |\mu_{k,i} - \mu_{k,i-1}|$. *By choosing* $\delta = \frac{1}{T}$ *and* $p = \sqrt{\frac{N_G K \log T}{T}}$, *the regret is upper-bounded by the following,*

$$\mathcal{O}\left(\left(\frac{\sum_{g\in G} N_g K_g \log T}{\Delta_{\min}} + \frac{\sqrt{N_G K T \log T}}{\left(\Delta_{\min}^{change}\right)^2} + N_{\mathbf{W}} K\right) \Delta_{\max}\right)$$

Our regret bound shows an improvement in comparison to the result of [37] in terms of the dependency of total restarts N_G, even though our algorithm does not require the prior knowledge of the causal graphs. In the absence of graph-changes, the respective contribution to the regret stems solely from the very first initialization, i.e., $N_{\mathbf{W}} = 1$.

5 Experimental Analysis

In this section, we evaluate the performance of our proposed decision-making policy using synthetic- and real-world datasets by comparing it with the following state-of-the-art combinatorial semi-bandit algorithms as benchmarks; **CTS** [17] is a Thompson sampling-based algorithm for stationary environments; **GLR-CUCB** [37] is a UCB-based algorithm for piecewise stationary environments. It employs a GLR change-point detector and uses a global restart strategy. We implemented the same algorithm with local restarts and group restarts, GLR-CUCB-Lo and GLR-CUCB-Gr, respectively; **CUCB-SW** [9] is an algorithm that uses a sliding window to follow the base arms' distribution changes while developing the corresponding UCB indices; **Orc-R** is PS-SEM-UCB-Gr with the Oracle-Restart. This algorithm is given the prior information w.r.t. all distribution change-points and it only restarts the groups where a change is detected. In addition, we implement the PS-SEM-UCB-Gl and PS-SEM-UCB-Lo that are working based on global restart and local restart strategy, respectively.

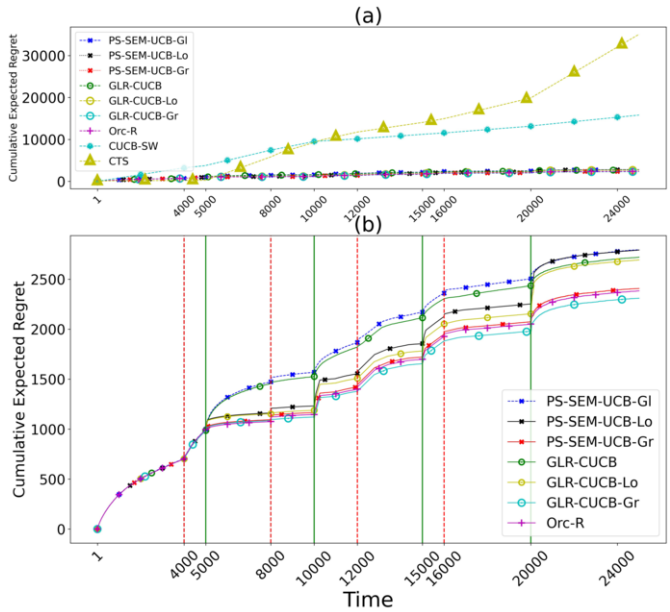

Figure 1. Cumulative Expected Regret.

All three algorithms *CTS*, *GLR-CUCB*, and *CUCB-SW* require access to the exact- or to an approximation oracle that solves the combinatorial optimization (11); that is, they need prior knowledge of the ground truth causal graph at any time. Such a strong assumption renders them inapplicable in the absence of such prior knowledge. For a fair comparison, we apply all benchmarks to the instantaneous rewards feedback vector $\mathbf{z}_t$ at each time t. We implemented the exact optimization oracles for *CTS*, *GLR-CUCB*, *GLR-CUCB-Gr*, *GLR-CUCB-Lo*, and *CUCB-SW*.

5.1 Synthetic Dataset

In the following, we describe the synthetic dataset used in the experiments. It has 4 graph-change-points and 4 distribution-change-points. For all different graph structures, we have $K = 18$ nodes. We draw the elements of the adjacency matrices $\mathbf{W}_t$ from a uniform distribution over $[0.1, 0.9]$. The edge density of the ground truth adjacency matrices is 0.15. The $K = 18$ arms are divided into 3 groups of 6 arms. We select $m = 4$ in this experiment and $T = 25000$. At each time t, the vector of instantaneous rewards $\mathbf{b}_t$ follows a multivariate normal distribution with the support in $[0, 1]^{18}$ and a spherical covariance matrix. In supplementary material, Figure 1 visualizes the expected values of base arms' rewards across time, and Figure 2 presents the visualization of optimal super arm across time. As shown in Section 2, the reward generation process follows the SEM in (3). All distribution-stationary-segments of the environment have the same lengths. The regularization parameter λ_1 is tuned by grid search over $[0.0001, 10000]$. We evaluate the estimated adjacency matrix at each time t by using the mean squared error defined as $\text{MSE} = \frac{1}{K^2} \left\| \mathbf{W}_t - \hat{\mathbf{W}}_t \right\|_{\text{F}}^2$. Figure 1-a shows the poor performance of *CUCB-SW*, and *CTS*, compared to other algorithms. In Figure 1-b, we highlight the differences in the performance of *Orc-R*, PS-SEM-UCB, *GLR-CUCB* under various restarting strate-

gies. One can observe the better performance of PS-SEM-UCB-Gr compared to *GLR-CUCB*. That happens although PS-SEM-UCB-Gr does not require prior knowledge of the distribution-change-points and the causal graphs. The effects of different restarting strategies can be observed as well. Global restarts adds to the regret significantly by restarting the entire set of base arms. On the opposite, local restart suffers from the delay on those breakpoints where the number of changed distributions is large. In Figure 1-b, each vertical green solid line represents the time of a distribution-change, and each vertical red dashed line represents a graph-change.

5.2 *Real-World Application*

In this section, we provide the results of applying our algorithm, to the Covid-19 outbreak dataset of daily new infected cases during the pandemic in different regions within Italy.[1] The goal is to find a subset of regions with the highest contribution to the spread of the virus in the country in a non-stationary period. We use the *overall reward* y_i for the *overall daily new cases* in region i. Besides, we use the *instantaneous reward* b_i for the *region-specific daily new cases* in region i. The data of the period from 3^{rd} July 2020, to 10^{th} October 2020 was used. We pre-process the dataset following [26]; nevertheless, we use a 14-day moving average instead of a 7-day moving average. Instead of the L^1-norm in (8), we use the Directed Total Variation (DTV) $\sum_{i,j\in\mathcal{K}} \mathbf{W}[i,j] \sum_{h=1,\dots,t} [\mathbf{Y}[i,h] - \mathbf{Y}[j,h]]^+$ regularizer [28], where $[y]^+ = \max\{y, 0\}$. Since the causal spread of the disease might create cycles, we allow cyclic graphs as the result of the optimization problem (8). Considering that the ground truth graphs are not available, we use a cross-validation technique to tune the regularization parameter λ_1. We split the data into 10 subsets of 10 consecutive days. In each subset, one day is chosen uniformly at random to be included in the validation set, while the remaining 9 days are added to the train set. We calculate the prediction error at each time t by $Error(t) = \frac{1}{K|\mathfrak{v}(t)|} \sum_{i\in\mathfrak{v}(t)} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_1$ where $\mathfrak{v}(t)$ is the validation set at time t with cardinality $|\mathfrak{v}(t)|$. Besides, $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ are respectively the validation data, and the corresponding predicted value using the estimated graph for day i. Figure 2 compares the ground truth overall daily new cases and the predicted total daily new cases using the estimated graph in 3 days of the Covid-19 outbreak in our validation data.[2] According to Figure 2, our algorithm estimates the data for each region efficiently. This helps the agent to find the optimal decision vector. Regarding that the benchmarks need the prior knowledge of the causal graph, this real-world application highlights the drawbacks of the benchmarks. Considering the impacts of geographical factors on Covid-19 cases [34], we divide the country into 4 clusters, using *graph-based clustering* library of *Python*, based on Euclidean distances between regional capitals. In Figure 3, we show the regions that PS-SEM-UCB-Gr selects over time. On each day, the selected regions are highlighted by dark rectangles. PS-SEM-UCB-Gr finds changes in the distribution of the region-specific daily new cases of different regions belonging to each group. Consequently, it restarts the UCB procedure for all the groups within the period $t = 58$ and $t = 79$. Due to space limitations, the details about the groupings and their change-detection times are mentioned in the supplementary. We see that selected subsets of regions before and after the restart of the algorithm are different due to newly calculated UCB indices after the restarts. This shows how the main contributors to the spread of the virus changed from one stationary segment to the next.

[1] https://github.com/pcm-dpc/COVID-19

[2] Due to space limitations, we use abbreviations for region names. Table 1 in supplementary material lists the abbreviations together with the original names of the regions.

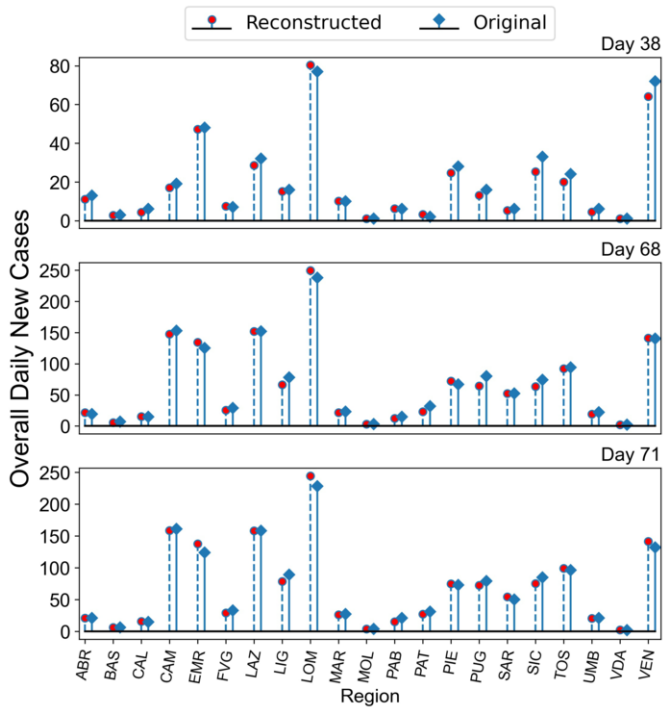

Figure 2. Original and reconstructed daily new cases.

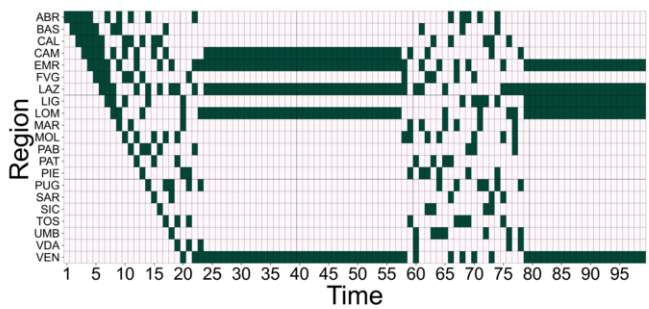

Figure 3. Selected regions on each day of the experiment.

6 Conclusion

In this paper, we developed a piecewise stationary combinatorial semi-bandit framework with causally related rewards. We developed a decision-making policy that follows distribution- and causal graph changes to adapt the decisions. We introduced a new alternative for the restarting process of bandit algorithms in structured environments under piecewise stationary settings. We proved that PS-SEM-UCB-Gr achieves a sublinear regret bound. The experiments showed the superior performance of PS-SEM-UCB-Gr compared to several state-of-the-art combinatorial algorithms. Our regret analysis clarifies the effects of global and local restarts as special cases of group restarts. It clarifies the importance of using relationships amongst base arms' distributions for the purpose of grouping of arms to minimize the regret incurred by the restarting process in group restarts. As for future research direction, we aim at studying our problem under the presence of noise in the SEM.

Acknowledgements

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645, and by Grant 16KISK035 from the German Federal Ministry of Education and Research (BMBF). We are grateful to Sofien Dhouib for fruitful discussions and comments.

References

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer, 'Finite-time analysis of the multiarmed bandit problem', *Machine learning*, **47**, 235–256, (2002).

[2] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire, 'The nonstochastic multiarmed bandit problem', *SIAM journal on computing*, **32**, 48–77, (2002).

[3] Maryam Aziz, Emilie Kaufmann, and Marie-Karelle Riviere, 'On multi-armed bandit designs for dose-finding clinical trials', *Journal of Machine Learning Research*, **22**, 4, (2021).

[4] Juan Andrés Bazerque, Brian Baingana, and Georgios B Giannakis, 'Identifiability of sparse structural equation models for directed and cyclic networks', in *2013 IEEE Global Conference on Signal and Information Processing*, pp. 839–842. IEEE, (2013).

[5] Omar Besbes, Yonatan Gur, and Assaf Zeevi, 'Stochastic multi-armed-bandit problem with non-stationary rewards', *Advances in neural information processing systems*, **27**, (2014).

[6] Lilian Besson and Emilie Kaufmann, 'The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits', *Proceedings of Machine Learning Research vol XX*, **1**, 35, (2019).

[7] Javier Borge-Holthoefer, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, David Francos, David Iniguez, María Pilar Pérez, Gonzalo Ruiz, et al., 'Structural and dynamical patterns on online social networks: the spanish may 15th movement as a case study', *PloS one*, **6**(8), e23883, (2011).

[8] Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie, 'Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit', in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 418–427. PMLR, (2019).

[9] Wei Chen, Liwei Wang, Haoyu Zhao, and Kai Zheng, 'Combinatorial semi-bandit in the non-stationary environment', in *Uncertainty in Artificial Intelligence*, pp. 865–875. PMLR, (2021).

[10] Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang, 'Combinatorial multi-armed bandit and its extension to probabilistically triggered arms', *The Journal of Machine Learning Research*, **17**, 1746–1778, (2016).

[11] Xiaotong Cheng and Setareh Maghsudi, 'Distributed consensus algorithm for decision-making in multi-agent multi-armed bandit', *arXiv preprint arXiv:2306.05998*, (2023).

[12] David Easley and Jon Kleinberg, *Networks, crowds, and markets: Reasoning about a highly connected world*, Cambridge university press, 2010.

[13] Aurélien Garivier and Eric Moulines, 'On upper-confidence bound policies for switching bandit problems', in *International Conference on Algorithmic Learning Theory*, pp. 174–188. Springer, (2011).

[14] Claudio Gentile, Shuai Li, and Giovanni Zappella, 'Online clustering of bandits', in *International Conference on Machine Learning*, pp. 757–765. PMLR, (2014).

[15] Georgios B Giannakis, Yanning Shen, and Georgios Vasileios Karanikolas, 'Topology identification and learning over graphs: Accounting for nonlinearities and dynamics', *Proceedings of the IEEE*, **106**, 787–807, (2018).

[16] Cédric Hartland, Nicolas Baskiotis, Sylvain Gelly, Michele Sebag, and Olivier Teytaud, 'Change point detection and meta-bandits for online learning in dynamic environments', in *CAp 2007: 9è Conférence francophone sur l'apprentissage automatique*, pp. 237–250, (2007).

[17] Alihan Huyuk and Cem Tekin, 'Analysis of thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms', in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1322–1330. PMLR, (2019).

[18] Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan, 'Cascading bandits: Learning to rank in the cascade model', in *International conference on machine learning*, pp. 767–776. PMLR, (2015).

[19] Finnian Lattimore, Tor Lattimore, and Mark D Reid, 'Causal bandits: Learning good interventions via causal inference', *Advances in Neural Information Processing Systems*, **29**, (2016).

[20] Lihong Li, Wei Chu, John Langford, and Robert E Schapire, 'A contextual-bandit approach to personalized news article recommendation', in *Proceedings of the 19th international conference on World wide web*, pp. 661–670, (2010).

[21] Shuai Li, Alexandros Karatzoglou, and Claudio Gentile, 'Collaborative filtering bandits', in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 539–548, (2016).

[22] Fang Liu, Joohyun Lee, and Ness Shroff, 'A change-detection based framework for piecewise-stationary multi-armed bandit problem', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, (2018).

[23] Setareh Maghsudi and Ekram Hossain, 'Multi-armed bandits with application to 5g small cells', *IEEE Wireless Communications*, **23**, 64–73, (2016).

[24] Joseph Mellor and Jonathan Shapiro, 'Thompson sampling in switching environments with bayesian online change detection', in *Artificial intelligence and statistics*, pp. 442–450. PMLR, (2013).

[25] Mark EJ Newman, 'Modularity and community structure in networks', *Proceedings of the national academy of sciences*, **103**(23), 8577–8582, (2006).

[26] Behzad Nourani-Koliji, Saeed Ghoorchian, and Setareh Maghsudi, 'Linear combinatorial semi-bandit with causally related rewards', in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. International Joint Conferences on Artificial Intelligence Organization, (2022).

[27] Herbert Robbins, 'Some aspects of the sequential design of experiments', *Bulletin of the American Mathematical Society*, **58**, 527–535, (1952).

[28] Stefania Sardellitti, Sergio Barbarossa, and Paolo Di Lorenzo, 'On the graph fourier transform for directed graphs', *IEEE Journal of Selected Topics in Signal Processing*, **11**(6), 796–811, (2017).

[29] Rasoul Shafipour, Santiago Segarra, Antonio G Marques, and Gonzalo Mateos, 'Identifying the topology of undirected networks from diffused non-stationary graph signals', *IEEE Open Journal of Signal Processing*, **2**, 171–189, (2021).

[30] Yanning Shen, Brian Baingana, and Georgios B Giannakis, 'Tensor decompositions for identifying directed graph topologies and tracking dynamic networks', *IEEE Transactions on Signal Processing*, **65**(14), 3675–3687, (2017).

[31] Dorina Thanou, Xiaowen Dong, Daniel Kressner, and Pascal Frossard, 'Learning heat diffusion graphs', *IEEE Transactions on Signal and Information Processing over Networks*, **3**(3), 484–499, (2017).

[32] Francesco Trovo, Stefano Paladino, Marcello Restelli, and Nicola Gatti, 'Sliding-window thompson sampling for non-stationary settings', *Journal of Artificial Intelligence Research*, **68**, 311–364, (2020).

[33] Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák, 'Spectral bandits for smooth graph functions', in *International Conference on Machine Learning*, pp. 46–54. PMLR, (2014).

[34] Danyang Wang, Xiaoxu Wu, Chenlu Li, Jiatong Han, and Jie Yin, 'The impact of geo-environmental factors on global covid-19 transmission: A review of evidence and methodology', *Science of the Total Environment*, 154182, (2022).

[35] Qinshi Wang and Wei Chen, 'Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications', *Advances in Neural Information Processing Systems*, **30**, (2017).

[36] Kaige Yang, Laura Toni, and Xiaowen Dong, 'Laplacian-regularized graph bandits: Algorithms and theoretical analysis', in *International Conference on Artificial Intelligence and Statistics*, pp. 3133–3143. PMLR, (2020).

[37] Huozhi Zhou, Lingda Wang, Lav Varshney, and Ee-Peng Lim, 'A near-optimal change-detection based algorithm for piecewise-stationary combinatorial semi-bandits', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6933–6940, (2020).