Linear Combinatorial Semi-Bandit with Causally Related Rewards

Behzad Nourani-Koliji[1*], **Saeed Ghoorchian**[2*] and **Setareh Maghsudi**[1]
[1]University of Tübingen
[2]Technical University of Berlin
behzad.nourani-koliji@uni-tuebingen.de saeed.ghoorchian@tu-berlin.de
setareh.maghsudi@uni-tuebingen.de

Abstract

In a sequential decision-making problem, having a structural dependency amongst the reward distributions associated with the arms makes it challenging to identify a subset of alternatives that guarantees the optimal collective outcome. Thus, besides individual actions' reward, learning the causal relations is essential to improve the decision-making strategy. To solve the two-fold learning problem described above, we develop the 'combinatorial semi-bandit framework with causally related rewards', where we model the causal relations by a directed graph in a stationary structural equation model. The nodal observation in the graph signal comprises the corresponding base arm's instantaneous reward and an additional term resulting from the causal influences of other base arms' rewards. The objective is to maximize the long-term average payoff, which is a linear function of the base arms' rewards and depends strongly on the network topology. To achieve this objective, we propose a policy that determines the causal relations by learning the network's topology and simultaneously exploits this knowledge to optimize the decision-making process. We establish a sublinear regret bound for the proposed algorithm. Numerical experiments using synthetic and real-world datasets demonstrate the superior performance of our proposed method compared to several benchmarks.

1 Introduction

In the seminal form of the Multi-Armed Bandit (MAB) problem, an agent selects an arm from a given set of arms at sequential rounds of decision-making. Upon selecting an arm, the agent receives a reward, which is drawn from the unknown reward distribution of that arm. The agent aims at maximizing the average reward over the gambling horizon [Robbins, 1952]. The MAB problem portrays the exploration-exploitation dilemma, where the agent decides between accumulating immediate reward and obtaining information that might result in larger reward only in the future [Maghsudi and Hossain, 2016]. To measure the performance of a strategy, one uses the notion of *regret*. It is the difference between the accumulated reward of the applied decision-making policy and that of the optimal policy in hindsight.

In a combinatorial semi-bandit setting [Chen *et al.*, 2013], at each round, the agent selects a subset of *base arms*. This subset is referred to as a *super arm*. She then observes the individual reward of each base arm that belongs to the selected super arm. Consequently, she accumulates the collective reward associated with the chosen super arm. The combinatorial MAB problem is challenging since the number of super arms is combinatorial in the number of base arms. Thus, conventional MAB algorithms such as [Auer *et al.*, 2002] are not appropriate for combinatorial problems as they result in suboptimal regret bounds. The aforementioned problem becomes significantly more difficult when there are causal dependencies amongst the reward distributions.

In some cases, it is possible to model the causal structure that affects the rewards [Lattimore *et al.*, 2016]. Therefore, exploiting the knowledge of this structure helps to deal with the aforementioned challenges. In our paper, we develop a novel combinatorial semi-bandit framework with causally related rewards, where we rely on Structural Equation Models (SEMs) [Kaplan, 2008] to model the causal relations. At each time of play, we see the instantaneous rewards of the chosen base arms as controlled stimulus to the causal system. Consequently, in our causal system, the solution to the decision-making problem is the choice over the exogenous input that maximizes the collected reward. We propose a decision-making policy to solve the aforementioned problem and prove that it achieves a sublinear regret bound in time. Our developed framework can be used to model various real-world problems, such as network data analysis of biological networks or financial markets. We apply our framework to analyze the development of Covid-19 in Italy. We show that our proposed policy is able to detect the regions that contribute the most to the spread of Covid-19 in the country.

Compared to previous works, our proposed framework does not require any prior knowledge of the structural dependencies. For example, in [Tang *et al.*, 2017], the authors exploit the prior knowledge of statistical structures to learn the best combinatorial strategy. At each decision-making round, the agent receives the reward of the selected super arm and some side rewards from the selected base arms' neighbors. In

[*]Equal Contribution

[Huyuk and Tekin, 2019] a Combinatorial Thompson Sampling (CTS) algorithm to solve a combinatorial semi-bandit problem with probabilistically triggered arms is proposed. The proposed algorithm has access to an oracle that determines the best decision at each round of play based on the already collected data. Similarly, the authors in [Chen *et al.*, 2016] study a setting where triggering super arms can probabilistically trigger other unchosen arms. They propose an Upper Confidence Bound (UCB)-based algorithm that uses an oracle to improve the decision-making process. In [Yu *et al.*, 2020], the authors formulate a combinatorial bandit problem where the agent has access to an influence diagram that represents the probabilistic dependencies in the system. The authors propose a Thompson sampling algorithm and its approximations to solve the formulated problem. Further, there are some works that study the underlying structure of the problem. For example, in [Toni and Frossard, 2018], the authors attempt to learn the structure of a combinatorial bandit problem. However, they do not assume any causal relations between rewards. Moreover, in [Sen *et al.*, 2017], the MAB framework is employed to identify the best soft intervention on a causal system while it is assumed that the causal graph is only partially unknown.

The rest of the paper is organized as follows. In Section 2, we formulate the structured combinatorial semi-bandit problem with causally related rewards. In Section 3, we introduce our proposed algorithm, namely SEM-UCB. Section 4 includes the theoretical analysis of the regret performance of SEM-UCB. Section 5 is dedicated to numerical evaluation. Section 6 concludes the paper.

2 Problem Formulation

Let $[N] = \{1, 2, \dots, N\}$ denote the set of *base arms*. $\mathbf{b}_t = [\mathbf{b}_t[1], \mathbf{b}_t[2], \dots, \mathbf{b}_t[N]] \in [0,1]^N$ represents the vector of *instantaneous rewards* of the base arms at time t. The instantaneous rewards of each base arm $i \in [N]$ are independent and identically distributed (i.i.d.) random variables drawn from an unknown probability distribution with mean $\boldsymbol{\beta}[i]$. We collect the mean rewards of all the base arms in the mean reward vector of $\boldsymbol{\beta} = [\boldsymbol{\beta}[1], \boldsymbol{\beta}[2], \dots, \boldsymbol{\beta}[N]]$.

We consider a causally structured combinatorial semi-bandit problem where an agent sequentially selects a subset of base arms over time. We refer to this subset as the *super arm*. More precisely, at each time t, the agent selects a *decision vector* $\mathbf{x}_t = [\mathbf{x}_t[1], \mathbf{x}_t[2], \dots, \mathbf{x}_t[N]] \in \{0,1\}^N$. If the agent selects the base arm i at time t, we have $\mathbf{x}_t[i] = 1$, otherwise $\mathbf{x}_t[i] = 0$. The agent observes the value of $\mathbf{b}_t[i]$ at time t only if $\mathbf{x}_t[i] = 1$. The agent is allowed to select at most s base arms at each time of play. Hence, we define the set of all feasible super arms as

$$\mathcal{X} = \left\{ \mathbf{x} \mid \mathbf{x} \in \{0,1\}^N \wedge \|\mathbf{x}\|_0 \leq s \right\}, \tag{1}$$

where $\|\cdot\|_0$ determines the number of non-zero elements in a vector. In our problem, the parameter s is pre-determined and is given to the agent.

We take advantage of a directed graph structure to model the causal relationships in the system. We consider an unknown stationary sparse Directed Acyclic Graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V}$ denotes the set of N vertices, i.e., $|\mathcal{V}| = N$, $\mathcal{E}$ is the edge set, and $\mathbf{A}$ denotes the weighted adjacency matrix. By $p \leq N-1$, we denote the length of the largest path in the graph $\mathcal{G}$. We assume that the reward generating processes in the bandit setting follow an error-free Structural Equation Model (SEM) ([Giannakis *et al.*, 2018], [Bazerque *et al.*, 2013]). The exogenous input vector and the endogenous output vector of the SEM at each time t are denoted by $\mathbf{z}_t = [\mathbf{z}_t[1], \mathbf{z}_t[2], \dots, \mathbf{z}_t[N]]$ and $\mathbf{y}_t = [\mathbf{y}_t[1], \mathbf{y}_t[2], \dots, \mathbf{y}_t[N]]$, respectively. At each time t, the exogenous input $\mathbf{z}_t$ represents the semi-bandit feedback in the decision-making problem. Formally,

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

where $\operatorname{diag}(\cdot)$ represents the diagonalization of its given input vector. Consequently, we define the elements of the endogenous output vector $\mathbf{y}_t$ as

$$\mathbf{y}_t[i] = \sum_{i \neq j} \mathbf{A}[i,j]\mathbf{y}_t[j] + \mathbf{F}[i,i]\mathbf{z}_t[i], \quad \forall i = 1, \dots, N, \tag{3}$$

where $\mathbf{F}$ is a diagonal matrix that captures the effects of the exogenous input vector $\mathbf{z}_t$. The SEM in Equation (3) implies that the output measurement $\mathbf{y}_t[i]$ depends on the single-hop neighbor measurements in addition to the exogenous input signal $\mathbf{z}_t[i]$. In our formulation, at each time t, the endogenous output $\mathbf{y}_t[i]$ represents the *overall reward* of the corresponding base arm $i \in [N]$. Therefore, at each time t, the overall reward of each base arm comprises two parts; one part directly results from its instantaneous reward, while the other part reflects the effect of causal influences of other base arms' overall rewards.

In Equation (3), the overall rewards are causally related. Thus, the adjacency matrix $\mathbf{A}$ represents the causal relationships between the overall rewards; accordingly, the element $\mathbf{A}[i,j]$ of the adjacency matrix $\mathbf{A}$ denotes the causal impact of the overall reward of base arm j on the overall reward of base arm i, and we have $\mathbf{A}[i,i] = 0$, $\forall i = 1, 2, \dots, N$. We assume that the agent is not aware of the causal relationships between the overall rewards. Hence, the adjacency matrix $\mathbf{A}$ is unknown a priori. In the following, we work with the matrix form of Equation (3), defined at time t as

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

In **Figure 1**, we illustrate an exemplary network consisting of N vertices and the underlying causal relations. Based on our problem formulation, the agent is able to observe both the exogenous input signal vector $\mathbf{z}_t$ and the endogenous output signal vector $\mathbf{y}_t$. As we see, there does not exist necessarily a causal relation between every pair of nodes.

By inserting (2) in (4) and solving for $\mathbf{y}_t$ we obtain

$$\mathbf{y}_t = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{F} \operatorname{diag}(\mathbf{b}_t)\mathbf{x}_t. \tag{5}$$

Finally, we define the *payoff* received by the agent upon choosing the decision vector $\mathbf{x}_t$ as

$$r(\mathbf{x}_t) = \mathbf{1}^\top \mathbf{y}_t = \mathbf{1}^\top (\mathbf{I} - \mathbf{A})^{-1}\mathbf{F} \operatorname{diag}(\mathbf{b}_t)\mathbf{x}_t, \tag{6}$$

where $\mathbf{1}$ is the N-dimensional vector of ones. Since the graph $\mathcal{G}$ is a DAG, it implies that with a proper indexing of the vertices, the adjacency matrix $\mathbf{A}$ is a strictly upper triangular matrix. This guarantees that the matrix $(\mathbf{I} - \mathbf{A})$ is invertible. In

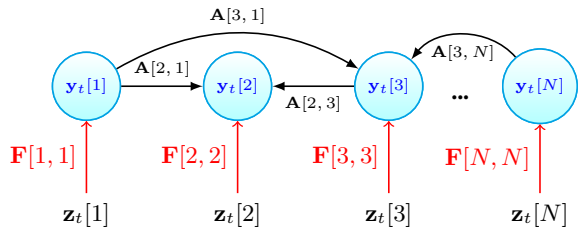

Figure 1: An exemplary illustration of a graph consisting of N vertices and their causal relations. The black directed edges represent the causal relationships amongst the vertices.

our problem, since the agent directly observes the exogenous input, we assume that the effects of $\mathbf{F}$ on the exogenous inputs are already integrated in the instantaneous rewards. Therefore, to simplify the notation and without loss of generality, we assume that $\mathbf{F} = \mathbf{I}$ in the following.

Given a decision vector $\mathbf{x}_t \in \mathcal{X}$, the expected payoff at time t is calculated as

$$\mu(\mathbf{x}_t) = \mathbb{E}\left[r(\mathbf{X})|\mathbf{X} = \mathbf{x}_t\right], \tag{7}$$

where the expectation is taken with respect to the randomness in the reward generating processes.

Ideally, the agent's goal is to maximize her total mean payoff over a time horizon T. Alternatively, the agent aims at minimizing the expected regret, defined as the difference between the expected accumulated payoff of an oracle that follows the optimal policy and that of the agent that follows the applied policy. Formally, the expected regret is defined as

$$\mathcal{R}_T(\mathcal{X}) = T\mu(\mathbf{x}^*) - \sum_{t=1}^{T} \mu(\mathbf{x}_t), \tag{8}$$

where $\mathbf{x}^* = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}}\ \mu(\mathbf{x})$ is the optimal decision vector, and $\mathbf{x}_t$ denotes the selected decision vector at time t under the applied policy.

Remark 1. *The definition of payoff in (6) implies that we are dealing with a linear combinatorial semi-bandit problem with causally related rewards. In general, due to the randomness in selection of the decision vector $\mathbf{x}_t$, the consecutive overall reward vectors $\mathbf{y}_t$ become non-identically distributed. In the following section, we propose our algorithm that is able to deal with such variables. This is an improvement over the previous methods, such as [Chen* et al.*, 2016] and [Huyuk and Tekin, 2019], that are not able to cope with our problem formulation, as they are specially designed to work with i.i.d. random variables.*

3 Decision-Making Strategy

In this section, we present our decision-making strategy to solve the problem described in Section 2. Our proposed policy consists of two learning components: (i) an online graph learning and (ii) an Upper Confidence Bound (UCB)-based reward learning. In the following, we describe each component separately and propose our algorithm, namely SEM-UCB.

3.1 Online Graph Learning

The payoff defined in (6) implies that the knowledge of $\mathbf{A}$ is necessary to select decision vectors that result in higher accumulated payoffs. Therefore, the agent aims at learning the matrix $\mathbf{A}$ to improve her decision-making process. To this end, we propose an online graph learning framework that uses the collected feedback, i.e., the collected exogenous input and endogenous output vectors, to estimate the ground truth matrix $\mathbf{A}$. In the following, we formalize the online graph learning framework.

At each time t, we collect the feedback up to the current time in $\mathbf{Z}_t = [\mathbf{z}_1 \dots \mathbf{z}_t]$ and $\mathbf{Y}_t = [\mathbf{y}_1 \dots \mathbf{y}_t]$. Therefore,

$$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_t + \mathbf{Z}_t. \tag{9}$$

We assume that the right indexing of the vertices is known prior to estimating the ground truth adjacency matrix. We use the collected feedback $\mathbf{Y}_t$ and $\mathbf{Z}_t$ as the input to a parametric graph learning algorithm ([Giannakis *et al.*, 2018], [Dong *et al.*, 2019]). More precisely, we use the following optimization problem to estimate the adjacency matrix at time t.

$$\begin{aligned} \hat{\mathbf{A}}_t = \underset{\mathbf{A}}{\operatorname{argmin}} \quad & \|\mathbf{Y}_t - \mathbf{A}\mathbf{Y}_t - \mathbf{Z}_t\|_2^2 + \mathrm{g}(\mathbf{A}) \\ \text{s.t.} \quad & \mathbf{A}[i,j] \geq 0, \quad \forall i,j \in [N], \\ & \mathbf{A}[i,j] = 0, \quad \forall i \geq j, \end{aligned} \tag{10}$$

where $\|\cdot\|_2$ represents the L^2-norm of matrices and $\mathrm{g}(\mathbf{A})$ is a regularization function that imposes sparsity over $\mathbf{A}$. In our numerical experiments, we work with different regularization functions to demonstrate the effectiveness of our proposed algorithm in different scenarios. As an example, we impose the sparsity property on the estimated matrix $\hat{\mathbf{A}}_t$ in (10) by defining $\mathrm{g}(\mathbf{A}) = \lambda \|\mathbf{A}\|_1$, where $\|\cdot\|_1$ is the L^1-norm of the matrices and λ is the regularization parameter. Our choices of regularization function guarantee that the optimization problem (10) is convex.

3.2 SEM-UCB Algorithm

We propose our decision-making policy in **Algorithm 1**. The key idea behind our algorithm is that it works with observations for each base arm, rather than the payoff observations for each super arm. As the same base arm can be observed while selecting different super arms, we can use the obtained information from selection of a super arm to improve our payoff estimation of other relevant super arms. This, combined with the fact that our algorithm simultaneously learns the causal relations, significantly improves the performance of our proposed algorithm and speed up the learning process.

For each base arm i, we define the empirical average of instantaneous rewards at time t as

$$\hat{\boldsymbol{\beta}}_t[i] = \frac{\sum_{\tau=1}^{t} \mathbf{b}_\tau[i] \mathbb{1}\left\{\mathbf{x}_\tau[i] = 1\right\}}{\mathbf{m}_t[i]}, \tag{11}$$

where $\mathbf{m}_t[i]$ denotes the number of times that the base arm i is observed up to time t. Formally,

$$\mathbf{m}_t[i] = \sum_{\tau=1}^{t} \mathbb{1}\left\{\mathbf{x}_\tau[i] = 1\right\}. \tag{12}$$

Algorithm 1 SEM-UCB: Structural Equation Model-Upper Confidence Bound

Input: Parameter s, initialization matrix $\mathbf{M}$.

for $t = 1, \ldots, N$ **do**
  Select column t of the initialization matrix $\mathbf{M}$ as the decision vector $\mathbf{x}_t$.
  Observe $\mathbf{z}_t$ and $\mathbf{y}_t$.
end for
for $t = N+1, \ldots, T$ **do**
  Solve (10) to obtain $\hat{\mathbf{A}}_{t-1}$.
  Calculate $\mathbf{E}_{t-1}[i]$ using (13), $\forall i \in [N]$.
  Select decision vector $\mathbf{x}_t$ that solves (14).
  Observe $\mathbf{z}_t$ and $\mathbf{y}_t$.
end for

The initialization phase of SEM-UCB algorithm follows a specific strategy to create a rich data that helps to learn the ground truth adjacency matrix. At each time t during the first N times of play, SEM-UCB picks the column t of an upper-triangular **initialization matrix** $\mathbf{M} \in \{0,1\}^{N\times N}$, where $\mathbf{M}$ is created as follows. All diagonal elements of $\mathbf{M}$ are equal to 1. As for the column i, if $i \leq s$, we set all elements above diagonal to 1. If $s+1 \leq i \leq N$, we select $s-1$ elements above diagonal uniformly at random and set them to 1. The remaining elements are set to 0.

After the initialization period, our proposed algorithm takes two steps at each time t to learn the causal relationships and the expected instantaneous rewards of the base arms. First, it uses the collected feedback $\mathbf{Y}_t$ and $\mathbf{Z}_t$ and solves the optimization problem (10) to obtain the estimated adjacency matrix. It then uses the reward observations to calculate the UCB index $\mathbf{E}_t[i]$ for each base arm i, defined as

$$\mathbf{E}_t[i] = \hat{\boldsymbol{\beta}}_t[i] + \sqrt{\frac{(s+1)\ln t}{\mathbf{m}_t[i]}}. \tag{13}$$

Afterward, the algorithm selects a decision vector $\mathbf{x}_t$ using the current estimate of the adjacency matrix and the developed UCB indices of the base arms. Let $\mathbf{E}_t = [\mathbf{E}_t[1], \mathbf{E}_t[2], \ldots, \mathbf{E}_t[N]]$. At time t, SEM-UCB selects $\mathbf{x}_t$ as

$$\begin{aligned} \mathbf{x}_t = \underset{\mathbf{x}\in\mathcal{X}}{\text{argmax}} \quad & \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{A}}_{t-1})^{-1} \operatorname{diag}(\mathbf{E}_{t-1})\mathbf{x} \\ \text{s.t.} \quad & \|\mathbf{x}\|_0 \leq s. \end{aligned} \tag{14}$$

Remark 2. *The initialization phase of our algorithm guarantees that all the base arms are pulled at least once and the matrix* $\mathbf{M}$ *is full rank. Consequently, the adjacency matrix* $\mathbf{A}$ *is uniquely identifiable from the collected feedback [Bazerque* et al.*, 2013].*

Remark 3. *Let* $\mathbf{c}^\top = \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{A}}_{t-1})^{-1} \operatorname{diag}(\mathbf{E}_{t-1})$. *Since all the elements of both matrices* $\mathbf{E}_{t-1}$ *and* $\hat{\mathbf{A}}_{t-1}$ *are non-negative, we have* $\mathbf{c}[i] > 0$, $\forall i \in [N]$. *Thus, the optimization problem (14) reduces to finding the* s*-biggest elements of* $\mathbf{c}$. *Therefore, (14) can be solved efficiently based on the choice of sorting algorithm used to order the elements of* $\mathbf{c}$.

The computational complexity of the SEM-UCB algorithm varies depending on the solver that is used to learn the graph. For example, if we use OSQP solver [Stellato et al.*, 2020], we achieve a computational complexity of order* $\mathcal{O}(N^4)$.

4 Theoretical Analysis

In this section, we prove an upper bound on the expected regret of SEM-UCB algorithm. We use the following definitions in our regret analysis. For any decision vector $\mathbf{x} \in \mathcal{X}$, let $\Delta(\mathbf{x}) = \mu(\mathbf{x}^*) - \mu(\mathbf{x})$. We define $\Delta_{\max} = \max_{\mathbf{x}:\mu(\mathbf{x})<\mu(\mathbf{x}^*)} \Delta(\mathbf{x})$ and $\Delta_{\min} = \min_{\mathbf{x}:\mu(\mathbf{x})<\mu(\mathbf{x}^*)} \Delta(\mathbf{x})$. Moreover, let $\mathbf{w}_t^\top = \mathbf{1}^\top(\mathbf{I} - \hat{\mathbf{A}}_t)^{-1}\operatorname{diag}(\mathbf{x}_{t+1})$. We define $w_{\max} = \max_t \max_i \mathbf{w}_t[i]$.

The following theorem states an upper bound on the expected regret of SEM-UCB.

Theorem 1. *The expected regret of SEM-UCB algorithm is upper bounded as*

$$\mathcal{R}_T(\mathcal{X}) \leq \Big[\frac{4w_{\max}^2 s^2(s+1)N\ln T}{\Delta_{\min}^2} + N + \frac{\pi^2}{3}s^p N\Big]\Delta_{\max}. \tag{15}$$

Proof. See Section 1 of supplementary material. ■

5 Experimental Analysis

In this section, we present experimental results to provide more insight on the usefulness of learning the causal relations for improving the decision-making process. We evaluate the performance of our algorithm on synthetic and real-world datasets by comparing it to standard benchmark algorithms.

Benchmarks. We compare SEM-UCB with state-of-the-art combinatorial semi-bandit algorithms that do not learn the causal structure of the problem. Specifically, we compare our algorithm with the following policies: (i) CUCB [Chen *et al.*, 2016] calculates a UCB index for each base arm at each time t and feeds them to an approximation oracle that outputs a super arm. (ii) DFL-CSR [Tang *et al.*, 2017] develops a UCB index for each base arm and selects a super arm at each time t based on a prior knowledge of a graph structure that shows the correlations among base arms. (iii) CTS [Huyuk and Tekin, 2019] employs Thompson sampling and uses an oracle to select a super arm at each time t. (iv) FTRL [Zimmert *et al.*, 2019] selects a super arm at each time t based on the method of Follow-the-Regularized-Leader. To be comparable, we apply these benchmarks on the vector of overall reward $\mathbf{y}_t$ at each time t. If a benchmark requires $\mathbf{y}_t$ to be in $[0,1]$, we feed the normalized version of $\mathbf{y}_t$ to the corresponding algorithm. Finally, in our experiments, we choose $s = 6$, meaning that the algorithms can choose 6 base arms at each time of play.

5.1 Synthetic Dataset

Our simulation setting is as follows. We first create a graph consisting of $N = 20$ nodes. The elements of the adjacency matrix $\mathbf{A}$ are drawn from a uniform distribution over $[0.4, 0.7]$. The edge density of the ground truth adjacency

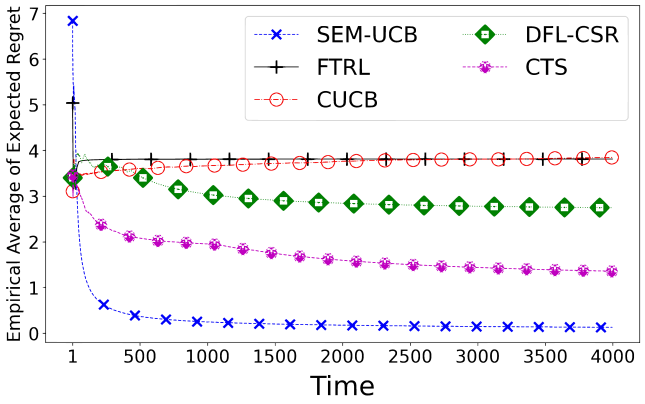

Figure 2: Time-averaged expected regret of different policies.

matrix is 0.15. At each time t, the vector of instantaneous rewards $\mathbf{b}_t$ is drawn from a multivariate normal distribution with the support in $[0, 1]^{20}$ and a spherical covariance matrix. As demonstrated in Section 2, we generate the vector of overall rewards according to the SEM in (3). We use $\mathrm{g}(\mathbf{A}) = \lambda \|\mathbf{A}\|_1$ as the regularization function in (10) when estimating the adjacency matrix $\mathbf{A}$. The regularization parameter λ is tuned by grid search over $[0.0001, 1000]$. We evaluate the estimated adjacency matrix at each time t by using the mean squared error defined as $\mathrm{MSE} = \frac{1}{N^2} \left\| \mathbf{A} - \hat{\mathbf{A}} \right\|_{\mathrm{F}}^2$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm.

Comparison with the benchmarks. We run the algorithms using the aforementioned synthetic data with $T = 4000$. In **Figure 2**, we depict the trend of time-averaged expected regret for each policy. As we see, SEM-UCB surpasses all other policies. This is due to the fact that SEM-UCB learns the network's topology and hence, it has a better knowledge of the causal relationships in the graph structure, unlike other policies that do not estimate the graph structure. As we see, the time-averaged expected regret of SEM-UCB tends to zero. This matches with our theoretical results in Section 4. Note that, the benchmark policies exhibit a sub-optimal regret performance as they have to deal with non-identically distributed random variables $\mathbf{y}_t$.

5.2 Covid-19 Dataset

We evaluate our proposed algorithm on the Covid-19 outbreak dataset of daily new infected cases during the pandemic in different regions within Italy.[1] The dataset fits in our framework as the daily new cases in each region results from the causal spread of Covid-19 among the regions in a country [Mastakouri and Schölkopf, 2020] and the region-specific characteristics [Guaitoli and Pancrazi, 2021]. As the regions differ in their regional characteristics, such as socio-economic and geographical characteristics, each region has a specific exposure risk of Covid-19 infection. To be consistent with our terminology in Section 2, at each time (day) t, we use the *overall reward* $\mathbf{y}_t[i]$ to refer to the *overall daily new cases* in region i and use the *instantaneous reward* $\mathbf{b}_t[i]$ to refer to the *region-specific daily new cases* in region i. Naturally, the overall daily new cases includes the region-specific daily new cases of Covid-19 infection.

Governments around the world strive to track the spread of Covid-19 and find the regions that are contributing the most to the total number of daily new cases in the country [Bridgwater and Bóta, 2021]. By the end of this experiment, we address this critical problem and highlight that our algorithm is capable of finding the optimal candidate regions for political interventions in order to contain the spread of a contagious disease such as Covid-19.

Data preparation. We focus on the recorded daily new cases from 10 August to 15 October, 2020, for $N = 21$ regions within Italy. The Covid-19 dataset only provides us with the overall daily new cases of each region. Hence, in order to apply our algorithm, we need to infer the distribution of region-specific daily new cases for each region. In the following, we describe this process and further pre-processing of the Covid-19 dataset.

According to [Bull, 2021], for the time period from 18 May to 3 June, 2020, all places for work and leisure activities were opened and travelling within regions was permitted while travelling between regions was forbidden. Consequently, during this period, there are no causal effects on the overall daily new cases of each region from other regions. In addition, according to google mobility data [Nouvellet *et al.*, 2021], during 4 weeks prior to 18 May the mobility was increasing within the regions while travel ban between the regions was still imposed. Hence, we use this expanded period to estimate the underlying distributions of the region-specific daily new cases using a kernel density estimation. Finally, considering that the daily recorded data noticeably fluctuates, a 7-day moving average was applied to the signals.

We create the region-specific daily new cases for each region by sampling from the estimated distributions. Below, we present the results of applying our algorithm on the pre-processed Covid-19 dataset. Since the data only contains the reported overall daily new cases for a limited time period, care should be exercised in interpreting the results. However, by providing more relevant data, our proposed framework helps towards more accurate detection of the regions that contribute the most to the development of Covid-19.

Learning the structural dependencies. Our algorithm learns the ground truth adjacency matrix $\mathbf{A}$ using (10). As for the choice of regularization function in (10), we employ Directed Total Variation (DTV) which is a novel application of the Graph Directed Variation (GDV) function [Sardellitti *et al.*, 2017]. DTV regularization function is defined as

$$\mathrm{g}(\mathbf{A}) = \lambda \sum_{i,j=1,\dots,N} \mathbf{A}[i,j] \sum_{k=1,\dots,t} \left[\mathbf{Y}[i,k] - \mathbf{Y}[j,k]\right]^+, \tag{16}$$

$$[y]^+ = \max\{y, 0\}. \tag{17}$$

The regularization function addresses the smoothness of the entire observations $\mathbf{Y}$ over the underlying directed graph. To be more realistic, since the causal spread of the disease might

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

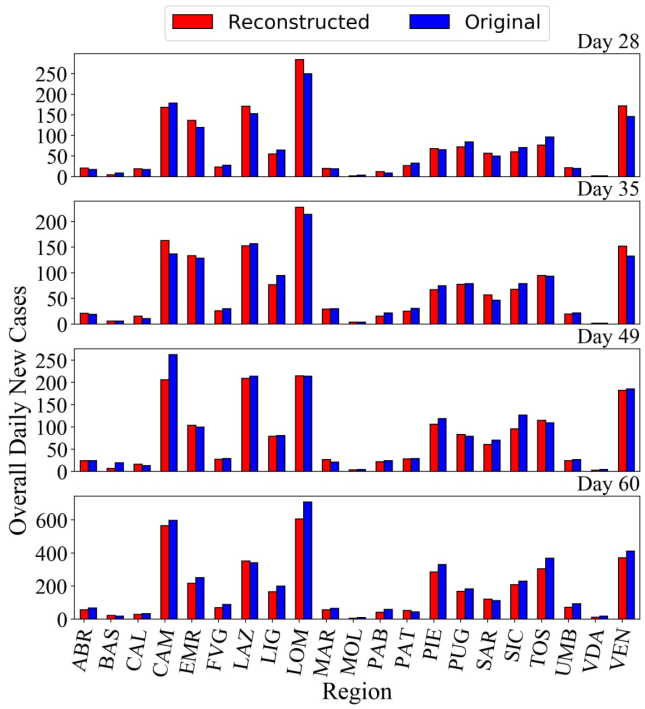

Figure 3: Original overall daily new cases and the corresponding predicted values for different days in the validation set.

create cycles, we additionally include cyclic graphs in the search space of the optimization problem (10).

We perform cross-validation technique to tune the regularization parameter λ. As mentioned before, we work on a limited time period with $T = 66$ days. Thus, we split the data into train and validation sets in 10:1 ratio. More specifically, we split the data into 6 subsets of 11 consecutive days. In each subset, one day is chosen uniformly at random to be included in the validation set while the remaining 10 days are added to the train set. We calculate the prediction error at each time t by

$$Error(t) = \frac{1}{NK(t)} \sum_{i \in \mathcal{K}(t)} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_1 , \qquad (18)$$

where $\mathcal{K}(t)$ is the validation set at time t with cardinality $K(t) = |\mathcal{K}(t)|$ and $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ are the validation data and the corresponding predicted value using the estimated graph for day i, respectively. **Figure 3** compares the ground truth overall daily new cases and the predicted overall daily new cases using the estimated graph on 4 different days of the Covid-19 outbreak in our validation data. Due to space limitation, we use abbreviations for region names. Table 1 in Section 2.1 of supplementary material lists the abbreviations together with the original names of the regions. We observe that our proposed framework is capable to estimate the data for each region efficiently, that helps the agent to improve its decision-making process in a real-world scenario.

Learning regions with the highest contribution. In **Figure 4**, we show the decision-making process of the agent over time by following the SEM-UCB policy. Dark rectangles represent the 6 selected regions at each day (time). Based on our framework, we represent the selected regions by our algorithm as those with biggest contributions to the development of Covid-19 during the time interval considered in our experiment. More specifically, we find the regions of Lombardia, Emilia-Romagna, Lazio, Veneto, Piemonte, and Liguria as the ones that contribute the most to the spread of Covid-19 during that period in Italy.

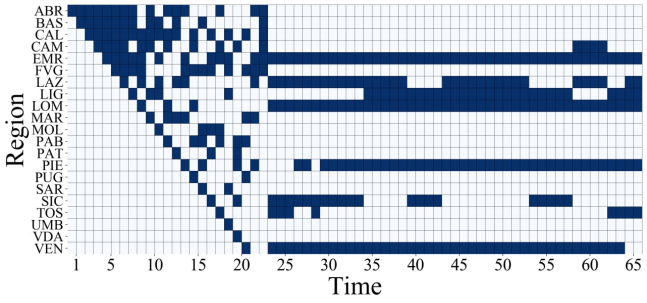

Figure 4: Selected regions on each day.

We emphasize that, due to the causal effects among the regions, contribution of each region to the spread of covid-19 differs from its overall daily cases of infection. Thus, the set of regions with the highest contribution does not necessarily equal to the set of regions with the highest total number of daily cases. This is a key aspect of our problem formulation that is addressed by SEM-UCB in **Figure 4**. We elaborate more on this fact in Section 2.3 of supplementary material.

6 Conclusion

In this paper, we developed a combinatorial semi-bandit framework with causally related rewards, where we modelled the causal relations by a directed graph in a structural equation model. We developed a decision-making policy, namely SEM-UCB, that learns the structural dependencies to improve the decision-making process. We proved that SEM-UCB achieves a sublinear regret bound in time. Our framework is applicable in a number of contexts such as network data analysis of biological networks or financial markets. We applied our method to analyze the development of Covid-19. The experiments showed that SEM-UCB outperforms several state-of-the-art combinatorial semi-bandit algorithms. Future research directions would be to extend the current framework to deal with piece-wise stationary environments where the causal graph and/or the expected instantaneous rewards of the base arms undergo abrupt changes over time.

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 01IS20051 from the German Federal Ministry of Education and Research (BMBF). We are grateful to Sergio Barbarossa and Sofien Dhouib for fruitful discussions and comments.

References

[Auer *et al.*, 2002] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[Bazerque *et al.*, 2013] 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*, pages 839–842. IEEE, 2013.

[Bridgwater and Bóta, 2021] Alexander Bridgwater and András Bóta. Identifying regions most likely to contribute to an epidemic outbreak in a human mobility network. In *2021 Swedish Artificial Intelligence Society Workshop (SAIS)*, pages 1–4. IEEE, 2021.

[Bull, 2021] Martin Bull. The italian government response to covid-19 and the making of a prime minister. *Contemporary Italian Politics*, pages 1–17, 2021.

[Chen *et al.*, 2013] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159. PMLR, 2013.

[Chen *et al.*, 2016] 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(1):1746–1778, 2016.

[Dong *et al.*, 2019] Xiaowen Dong, Dorina Thanou, Michael Rabbat, and Pascal Frossard. Learning graphs from data: A signal representation perspective. *IEEE Signal Processing Magazine*, 36(3):44–63, 2019.

[Giannakis *et al.*, 2018] 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(5):787–807, 2018.

[Guaitoli and Pancrazi, 2021] Gabriele Guaitoli and Roberto Pancrazi. Covid-19: Regional policies and local infection risk: Evidence from italy with a modelling study. *The Lancet Regional Health-Europe*, 8:100169, 2021.

[Huyuk and Tekin, 2019] 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*, pages 1322–1330. PMLR, 2019.

[Kaplan, 2008] David Kaplan. *Structural equation modeling: Foundations and extensions*, volume 10. Sage Publications, 2008.

[Lattimore *et al.*, 2016] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: learning good interventions via causal inference. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1189–1197, 2016.

[Maghsudi and Hossain, 2016] Setareh Maghsudi and Ekram Hossain. Multi-armed bandits with application to 5g small cells. *IEEE Wireless Communications*, 23(3):64–73, 2016.

[Mastakouri and Schölkopf, 2020] Atalanti Mastakouri and Bernhard Schölkopf. Causal analysis of covid-19 spread in germany. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3153–3163. Curran Associates, Inc., 2020.

[Nouvellet *et al.*, 2021] Pierre Nouvellet, Sangeeta Bhatia, Anne Cori, Kylie EC Ainslie, Marc Baguelin, Samir Bhatt, Adhiratha Boonyasiri, Nicholas F Brazeau, Lorenzo Cattarino, Laura V Cooper, et al. Reduction in mobility and covid-19 transmission. *Nature communications*, 12(1):1–9, 2021.

[Robbins, 1952] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[Sardellitti *et al.*, 2017] 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.

[Sen *et al.*, 2017] Rajat Sen, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Identifying best interventions through online importance sampling. In *International Conference on Machine Learning*, pages 3057–3066. PMLR, 2017.

[Stellato *et al.*, 2020] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672, 2020.

[Tang *et al.*, 2017] Shaojie Tang, Yaqin Zhou, Kai Han, Zhao Zhang, Jing Yuan, and Weili Wu. Networked stochastic multi-armed bandits with combinatorial strategies. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pages 786–793. IEEE, 2017.

[Toni and Frossard, 2018] Laura Toni and Pascal Frossard. Spectral mab for unknown graph processes. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 116–120. IEEE, 2018.

[Yu *et al.*, 2020] Tong Yu, Branislav Kveton, Zheng Wen, Ruiyi Zhang, and Ole J Mengshoel. Graphical models meet bandits: A variational thompson sampling approach. In *International Conference on Machine Learning*, pages 10902–10912. PMLR, 2020.

[Zimmert *et al.*, 2019] Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pages 7683–7692. PMLR, 2019.