Causal Bandits with Propagating Inference

Akihiro Yabe [1] **Daisuke Hatano** [2] **Hanna Sumita** [3] **Shinji Ito** [1] **Naonori Kakimura** [4] **Takuro Fukunaga** [2] **Ken-ichi Kawarabayashi** [5]

Abstract

Bandit is a framework for designing sequential experiments, where a learner selects an arm $A \in \mathcal{A}$ and obtains an observation corresponding to A in each experiment. Theoretically, the tight regret lower-bound for the general bandit is polynomial with respect to the number of arms $|\mathcal{A}|$, and thus, to overcome this bound, the bandit problem with side-information is often considered. Recently, a bandit framework over a causal graph was introduced, where the structure of the causal graph is available as side-information and the arms are identified with interventions on the causal graph. Existing algorithms for causal bandit overcame the $\Omega(\sqrt{|\mathcal{A}|/T})$ simple-regret lower-bound; however, their algorithms work only when the interventions $\mathcal{A}$ are localized around a single node (i.e., an intervention propagates only to its neighbors). We then propose a novel causal bandit algorithm for an arbitrary set of interventions, which can propagate throughout the causal graph. We also show that it achieves $O(\sqrt{\gamma^* \log(|\mathcal{A}|T)/T})$ regret bound, where γ^* is determined by using a causal graph structure. In particular, if the maximum in-degree of the causal graph is a constant, then $\gamma^* = O(N^2)$, where N is the number of nodes.

1. Introduction

Multi-armed bandit has been widely recognized as a standard framework for modeling online learning with a limited number of observations. In each round in the bandit problem, a learner chooses an arm A from given candidates $\mathcal{A}$, and obtains a corresponding observation. Since observation is limited, the learner must adopt an efficient strategy for exploring the optimal arm $A^* \in \mathcal{A}$. The efficiency of the strategy is measured by regret, and the theoretically tight lower-bound is $O(\sqrt{|\mathcal{A}|})$ with respect to the number of arms $|\mathcal{A}|$ in the general multi-armed bandit setting. Thus, in order to improve the above lower bound, one requires additional information for the bandit setting. For example, contextual bandit (Agarwal et al., 2014; Auer et al., 2002) is a well-known class of bandit problems with side information on domain-expert knowledge. For this setting, there is a logarithmic regret bound $O(\sqrt{\log |\mathcal{A}|})$ with respect to the number of arms. In this paper, we also achieve $O(\sqrt{\log |\mathcal{A}|})$ regret bound for a novel class of bandit problems with side information. To this end, let us introduce our bandit setting in detail.

Causal graph (Pearl, 2009) is a well-known tool for modeling a variety of real problems, including computational advertising (Bottou et al., 2013), genetics (Meinshausen et al., 2016), agriculture (Splawa-Neyman et al., 1990), and marketing (Kim et al., 2008). Based on causal graph discovery studies (Eberhardt et al., 2005; Hauser & Bühlmann, 2014; Hu et al., 2014; Shanmugam et al., 2015), Lattimore et al. (2016) recently introduced the causal bandit framework. They consider the problem of finding the best intervention which causes desirable propagation of a probabilistic distribution over a given causal graph with a limited number of experiments T. In this setting, the arms are identified as interventions $\mathcal{A}$ on the causal graph. A set of binary random variables $V_1, V_2, \ldots, V_N$ is associated with nodes $v_1, v_2, \ldots, v_N$ of the causal graph. At each round of an experiment, a learner selects an intervention $A \in \mathcal{A} \subseteq \{0, 1, *\}^N$ which enforces a realization of a variable V_i to A_i when $A_i \in \{0, 1\}$. The effect of the intervention then propagates throughout the causal graph through the edges, and a realization $\omega \in \{0, 1\}^N$ over all nodes is observed after propagation. The goal of the causal bandit problem is to control the realization of a target variable V_N with an optimal intervention.

Figure 1 is an illustrative example of the causal bandit problem. In the figure, the four nodes on the right represent a consumer decision-making model in e-commerce borrowed from (Kim et al., 2008). This model assumes that customers make a decision to purchase based on their perceived risk in an online transition (e.g., defective product), the con-

[1]NEC Corporation, Japan [2]RIKEN AIP, Japan [3]Tokyo Metropolitan University, Japan [4]Keio University, Japan [5]National Institute of Informatics, Japan. Correspondence to: Akihiro Yabe <a-yabe@cq.jp.nec.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

sumer's trust of a web vendor, and the perceived benefit in e-commerce (e.g., increased convenience). Consumer trust influences perceived risk. Here, we consider controlling customer's behavior by two kinds of advertising that correspond to adding two nodes (Ad A and Ad B) to be intervened into the model. Ad A can change only the reliability of a website, that is, it can influence the decision of customers in an indirect way through the middle nodes. In contrast, Ad B can change the perceived benefit. The aim is to increase the number of purchases by consumers through choosing an effective advertisement. This is indeed a bandit problem over a causal graph.

The work in (Lattimore et al., 2016) considered the causal bandit problem to minimize simple regret and offered an improved regret bound over the aforementioned tight lower-bound $\Omega(\sqrt{|\mathcal{A}|/T})$ (Audibert & Bubeck, 2010)[Theorem 4] for the general bandit setting (Audibert & Bubeck, 2010; Gabillon et al., 2012). Sen et al. (2017) extended this study by incorporating a smooth intervention, and they provided a new regret bound parameterized by the performance gap between the optimal and sub-optimal arms. This parameterized bound comes from the technique developed for the general multi-armed bandit problem (Audibert & Bubeck, 2010). These analyses, however, only work for a special class of interventions with known true parameters. Indeed, they only consider localized interventions.

Main contribution This paper proposes the first algorithm for the causal bandit problem with an arbitrary set of interventions (which can propagate throughout the causal graph), with a theoretically guaranteed simple regret bound. The bound is $O(\sqrt{\gamma^* \log(|\mathcal{A}|T)/T})$, where γ^* is a parameter bounded on the basis of the graph structure. In particular, $\gamma^* = O(N^2)$ if the maximum in-degree of the causal graph is bounded by a constant, where N is the number of nodes.

The major difficulty in dealing with an arbitrary intervention comes from accumulation and propagation of estimation error. Existing studies consider interventions that only affect the parents $\mathcal{P}_k$ of a single node V_k. To estimate the relationship between $\mathcal{P}_k$ and V_k in this setting, we could apply an efficient importance sampling algorithm (Bottou et al., 2013; Lattimore et al., 2016). On the other hand, when we intervene an arbitrary node, it can affect the probabilistic propagation mechanism in any part of the causal graph. Hence, we cannot directly control the realization of intermediate nodes when designing efficient experiments.

The proposed algorithm consists of two steps. First, the preprocessing step is devoted to estimating parameters for designing efficient experiments used in the main step. More precisely, we focus on estimation of parameters with bounded *relative error*. By truncating small parameters that are negligible but tend to have large relative error, we manage to avoid accumulation of estimation error. In the main step, we apply an importance sampling approach introduced in (Lattimore et al., 2016; Sen et al., 2017) on the basis of estimated parameters with a guaranteed relative error. This step allows us to estimate parameters with bounded *absolute error*, which results in the desired regret bound.

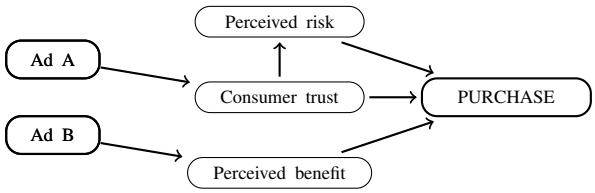

Figure 1. Simple example of a causal graph.

Owing to space limitations, all the proofs are omitted, where they can be found in the full version of this paper (Yabe et al., 2018).

Related studies Minimizing simple regret in bandit problems is called the best-arm identification (Gabillon et al., 2012; Kaufmann et al., 2016) or pure exploration (Bottou et al., 2013) problem, and it has been extensively studied in the machine learning research community. The inference of a causal graph structure is also well-studied, which can be classified into causal graph discovery and causal inference: Causal graph discovery (Eberhardt et al., 2005; Hauser & Bühlmann, 2014; Hu et al., 2014; Shanmugam et al., 2015) considers efficient experiments for determining the structure of causal graph, while causal inference (Mooij et al., 2016; Pearl, 2009; Shimizu et al., 2011; Spirtes & Glymour, 1991) challenges one to determine the graph structure only from historical data without additional experiments. The causal bandit problem designs experiments without using historical data, which is rather compatible with causal graph discovery studies.

2. Causal bandit problem

This section introduces the causal bandit problem proposed by (Lattimore et al., 2016).

Let $G = (\mathcal{V}, E)$ be a directed acyclic graph (DAG) with a node set $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ and a (directed) edge set E. Let (v_i, v_j) denote an edge from v_i to v_j. Without loss of generality, we suppose that the nodes in $\mathcal{V}$ are topologically sorted so that no edge from v_i to v_j exists if $i \geq j$. For each $n = 1, \ldots, N$, let $\mathcal{P}_n$ denote the index set of the parents of v_n, i.e., $\mathcal{P}_n = \{i \in \{1, \ldots, n-1\} \colon (v_i, v_n) \in E\}$. We then define $\overline{\mathcal{P}_n} = \mathcal{P}_n \cup \{n\}$.

Each node $v_n \in \mathcal{V}$ is associated with a random variable V_n, which takes a value in $\{0, 1\}$. The distribution of V_n is then influenced by the variables associated with the parents

of v_n (unless V_n is intervened, as described below). For each $\pi \in \{0,1\}^{\overline{\mathcal{P}_n}}$, the parameter $\alpha_n(\pi)$ defined below characterizes the distribution of V_n given the realizations of its parents:

$$\alpha_n(\pi) := \mathrm{Prob}\left(V_n = \pi_n \;\middle|\; \begin{array}{l} V_i = \pi_i \text{ for all } i \in \mathcal{P}_n, \\ v_n \text{ is not intervened} \end{array}\right).$$

That is to say, if the parents v_i for $i \in \mathcal{P}_n$ are realized as π_i, then $V_n = \pi_n$ with probability $\alpha_n(\pi)$, and $V_n = 1 - \pi_n$ with probability $1 - \alpha_n(\pi)$.

Together with a DAG, we are also given a set $\mathcal{A}$ of interventions. Each intervention is identified with a vector $A \in \{*, 0, 1\}^N$, where $A_n \neq *$ implies that V_n is intervened and that the realization of V_n is fixed as A_n. Let $\pi \in \{0,1\}^{\overline{\mathcal{P}_n}}$. Given an intervention $A \in \mathcal{A}$ and realizations π_i over the parents $i \in \mathcal{P}_n$, the probability that $V_n = \pi_n$ holds is then determined as follows:

$$\begin{aligned} &\mathrm{Prob}\left(V_n = \pi_n \mid V_i = \pi_i \text{ for all } i \in \mathcal{P}_n, \mathrm{do}(A)\right) \\ &= \begin{cases} \alpha_n(\pi) & \text{if } A_n = *, \\ 1 & \text{if } A_n = \pi_n, \\ 0 & \text{if } A_n = 1 - \pi_n. \end{cases} \end{aligned}$$

This equality together with the adjacency of the causal graph G completely determines the joint distribution over the variables $V_1, V_2, \ldots, V_N$, under an arbitrary intervention $A \in \mathcal{A}$.

In the causal bandit problem, we are given a DAG $G = (\mathcal{V}, E)$ and a set $\mathcal{A}$ of interventions. However, the parameters α_n $(n = 1, \ldots, N)$ are not known. Our ideal goal is then to find an intervention $A^* \in \mathcal{A}$ that maximizes the probability $\mu(A^*)$ of realizing $V_N = 1$, where $\mu : \mathcal{A} \to [0,1]$ is defined by

$$\mu(A) := \mathrm{Prob}(V_N = 1 \mid \mathrm{do}(A))$$

for each $A \in \mathcal{A}$.

For this purpose, we discuss the following algorithms. First, they estimate $\mu(A)$ $(A \in \mathcal{A})$ from T experimental trials. Each experiment consists of the application of an intervention and the observation of a realization $\pi \in \{0,1\}^N$ over all nodes. Let $\hat{\mu}(A)$ denote the estimate of $\mu(A)$. Second, the algorithm selects the intervention $\hat{A}$ that maximizes $\hat{\mu}$. We evaluate the efficiency of such an algorithm with the simple regret R_T defined as follows:

$$R_T = \mu(A^*) - E[\mu(\hat{A})].$$

Note that, even if an algorithm is deterministic, $\hat{A}$ includes stochasticity since the observations obtained in each experiment are produced by a stochastic process.

In this paper, we assume that $N \geq 3$ and $T \geq 2$ for ease of technical discussion.

3. Proposed Algorithm

We propose an algorithm for the causal bandit problem, and present a regret bound of the proposed algorithm in this section. Let $C_n = 2^{|\mathcal{P}_n|}$ for each $n = 1, \ldots, N$, and $C = \sum_{n=1}^{N} C_n$. For $S \subseteq S' \subseteq [1, N]$ and $\pi \in \{0,1\}^{S'}$, let π_S denote the restriction of π onto S.

3.1. Outline of the proposed algorithm

Recall that the purpose of the causal bandit problem is to identify an intervention A^* that maximizes $\mu(A^*)$. This task is trivial if α_n is known for all $n = 1, \ldots, N$, because $\mu(A)$ can then be calculated for all $A \in \mathcal{A}$. Let $B(A) = \{\pi' \in \{0,1\}^N \mid \pi'_i = A_i \text{ if } A_i \neq *, \pi'_N = 1\}$, and for $n \in [1, N]$, let $I_{n,A}$ denote the set of nodes in $[1, n]$ which are not intervened by A; $I_{n,A} := \{m \in [1, n] \mid A_m = *\}$. $\mu(A)$ can then be represented as

$$\mu(A) = \sum_{\pi \in B(A)} \prod_{n \in I_{N,A}} \alpha_n(\pi_{\overline{\mathcal{P}_n}}).$$

Therefore, for computing μ approximately, our algorithm estimates α_n $(n = 1, \ldots, N)$.

In order to estimate α_n efficiently, we are required to manipulate the random variables associated with the parents of v_n. More concretely, to estimate $\alpha_n(\overline{\pi})$ for $\overline{\pi} \in \overline{\mathcal{P}_n}$, we require samples with realization $\omega \in \{0,1\}^N$ satisfying $\overline{\pi}_i = \omega_i$ over the parents $i \in \mathcal{P}_n$ of v_n. For $n = 1, 2, \ldots, N$, $\pi \in \{0,1\}^{\mathcal{P}_n}$, and $A \in \mathcal{A}$, we thus introduce the additional quantities $\beta_n(\pi, A)$ that denote the probability of realizing ω with $\omega_{\mathcal{P}_n} = \pi$ under a given intervention A. More precisely, we define

$$\begin{aligned} &\beta_n(\pi, A) \\ &:= \begin{cases} \mathrm{Prob}(V_m = \pi_m, \forall m \in \mathcal{P}_n \mid \mathrm{do}(A)) & \text{if } A_n = *, \\ 0 \quad \text{otherwise.} \end{cases} \end{aligned}$$

Our algorithm consists of two phases. The first phase estimates β_n $(n = 1, \ldots, N)$, and the second phase estimates α_n $(n = 1, \ldots, N)$. The algorithm requires $T/3$ experiments in the first phase, and $2T/3$ experiments in the second phase. In the rest of this section, we first explain those phases and present a regret bound on the algorithm.

3.2. First Phase: Estimation of β

Here, we introduce the estimation phase of β_n for all $n = 1, \ldots, N$. The pseudo-code of this phase is described in Algorithm 1. Algorithm 1 requires a non-negative number λ as a parameter, which will be set to C^3/N. We perform $T/3$ experiments in this phase.

Before explaining the details of Algorithm 1, we note that β_n can be calculated from $\alpha_1, \ldots, \alpha_{n-1}$. For $\pi \in \{0,1\}^{\mathcal{P}_n}$, let $B_n(\pi, A) := \{\pi' \in \{0,1\}^{n-1} \mid \pi'_i = \pi_i \text{ if } i \in$

Algorithm 1 Estimation of β

Require: λ
Ensure: $\hat{\beta}_n$ $(n = 1, \dots, N)$ and D
$G_n \leftarrow \emptyset$ for $n = 1, 2, \dots, N$
for $n = 1, 2, \dots, N$ **do**
  for $\pi \in \{0,1\}^{\mathcal{P}_n}$ **do**
    Calculate $\hat{\beta}_n(\pi, A)$ for each $A \in \mathcal{A}$ by (2)
    Calculate $\hat{A}_{n,\pi} = \mathrm{argmax}_{A \in \mathcal{A}} \hat{\beta}_n(\pi, A)$
    $t_n(\pi) \leftarrow 0$ and $\overline{t_n}(\pi) \leftarrow 0$
    for $j = 1, \dots, T/(3C)$ **do**
      Conduct an experiment with $\hat{A}_{n,\pi}$ and let $\omega \in \{0,1\}^N$ be the obtained result
      $t_n(\pi) \leftarrow t_n(\pi) + 1$ if $\omega_i = \pi_i$ for all $i \in \mathcal{P}_n$
      $\overline{t_n}(\pi) \leftarrow \overline{t_n}(\pi) + 1$ if $\omega_i = \pi_i$ for all $i \in \mathcal{P}_n$ and $\omega_n = 1$
    end for
    for $k = 0, 1$ **do**
      Extend π to $\pi' \in \{0,1\}^{\overline{\mathcal{P}_n}}$ with $\pi'_n = k$
      Compute $\check{\alpha}'_n(\pi')$ by (3)
      If (4) holds, then $G_n \leftarrow G_n \cup \{\pi'\}$
      Compute $\check{\alpha}_n(\pi')$ by (5)
    end for
  end for
end for
Compute H_n and D_n $(n = 1, 2, \dots, N)$ by (6) and (7)
return $\hat{\beta}_n$ $(n = 1, 2, \dots, N)$ and $D = \{D_n \mid n = 1, 2, \dots, N\}$

Algorithm 2 Estimation of α

Require: $\hat{\beta}_n$ $(n = 1, \dots, N)$ and D
Ensure: $\hat{\alpha}_n$ $(n = 1, \dots, N)$
for $n = 1, 2, \dots, N$ and each $\pi \in \{0,1\}^{\mathcal{P}_n}$ **do**
  $t'_n(\pi) \leftarrow 0$ and $\overline{t'_n}(\pi) \leftarrow 0$
  Calculate $\hat{A}_{n,\pi} := \mathrm{argmax}_{A \in \mathcal{A}} \hat{\beta}_n(\pi, A)$
  for $j = 1, \dots, T/(3C)$ **do**
    Conduct an experiment with $\hat{A}_{n,\pi}$ and let $\omega \in \{0,1\}^N$ be the obtained result
    for $m = 1, \dots, N$ with $(\hat{A}_{n,\pi})_m = *$ **do**
      $t'_m(\omega_{\mathcal{P}_m}) \leftarrow t'_m(\omega_{\mathcal{P}_m}) + 1$
      $\overline{t'_m}(\omega_{\mathcal{P}_m}) \leftarrow \overline{t'_m}(\omega_{\mathcal{P}_m}) + 1$ if $\omega_m = 1$
    end for
  end for
end for
Compute an optimal solution $\hat{\eta}$ for (8)
for $t = 1, 2, \dots, T/3$ **do**
  Sample A_t from $\mathcal{U}(\hat{\eta})$
  Conduct experiment with A_t and let $\omega \in \{0,1\}^N$ be the obtained realization
  for $n = 1, \dots, N$ with $A_n = *$ **do**
    $t'_n(\omega_{\mathcal{P}_n}) \leftarrow t'_n(\omega_{\mathcal{P}_n}) + 1$
    $\overline{t'_n}(\omega_{\mathcal{P}_n}) \leftarrow \overline{t'_n}(\omega_{\mathcal{P}_n}) + 1$ if $\omega_n = 1$
  end for
end for
for $n = 1, 2, \dots, N$ and $\pi \in \{0,1\}^{\overline{\mathcal{P}_n}}$ **do**
  Compute $\hat{\alpha}'_n(\pi)$ by (9) and $\hat{\alpha}_n(\pi)$ by (10)
end for
return $\hat{\alpha}_n$.

Algorithm 3 Causal Bandit

Apply Algorithm 1 with $\lambda = C^3/N$ to obtain $\hat{\beta}_n$ $(n = 1, \dots, N)$ and D
Apply Algorithm 2 to obtain $\hat{\alpha}_n$ $(n = 1, \dots, N)$
Calculate $\hat{\mu}(A)$ for each $A \in \mathcal{A}$ by (11)
return $\hat{A} := \mathrm{argmax}_{A \in \mathcal{A}} \hat{\mu}(A)$

$\mathcal{P}_n, \pi'_i = A_i$ if $A_i \neq *\}$ denote the set of realizations over $V_1, V_2, \dots, V_{n-1}$ that is consistent with the realization π over $\mathcal{P}_n$ and the intervention A. If $A_n = *$, then $\beta_n(\pi, A)$ is described as

$$\beta_n(\pi, A) = \sum_{\pi' \in B_n(\pi, A)} \prod_{m \in I_{n-1,A}} \alpha_m(\pi'_{\mathcal{P}_m}). \tag{1}$$

Algorithm 1 consists of N iterations. The n-th iteration computes the following objects:

- an estimate $\hat{\beta}_n$ of β_n,
- $\hat{A}_{n,\pi} \in \mathcal{A}$ for each $\pi \in \{0,1\}^{\mathcal{P}_n}$,
- an estimate $\check{\alpha}_n$ of α_n, and
- $G_n \subseteq \{0,1\}^{\overline{\mathcal{P}_n}}$.

We remark that $\check{\alpha}_n$ in Algorithm 1 are used only for computing an estimate $\hat{\beta}_n$ and are not used for estimating μ. An estimate of α_n is computed in the next phase of our algorithm.

At the beginning of the n-th iteration, we compute $\hat{\beta}_n(\pi, A)$ for each $\pi \in \{0,1\}^{\mathcal{P}_n}$ and $A \in \mathcal{A}$ by (1) substituting $\check{\alpha}_m$ for α_m;

$$\hat{\beta}_n(\pi, A) = \sum_{\pi' \in B_n(\pi, A)} \prod_{m \in I_{n-1,A}} \check{\alpha}_m(\pi'_{\mathcal{P}_n}). \tag{2}$$

Let us confirm that this $\hat{\beta}_n(\pi, A)$ can be computed if $\check{\alpha}_m$ $(m = 1, \dots, n-1)$ are available.

For each $\pi \in \{0,1\}^{\mathcal{P}_n}$, then, we identify an intervention $\hat{A}_{n,\pi}$ that attains $\max_{A \in \mathcal{A}} \hat{\beta}_n(\pi, A)$. Using $\hat{A}_{n,\pi}$, we compute $\check{\alpha}_n(\overline{\pi})$ as follows, where $\overline{\pi}$ is an extension of π onto $\{0,1\}^{\overline{\mathcal{P}_n}}$. We conduct $T/(3C)$ experiments with $\hat{A}_{n,\pi}$. Let $t_n(\pi)$ be the number of experiments in those $T/(3C)$ experiments in which the obtained realization $\omega \in \{0,1\}^N$ satisfies $\omega_i = \pi_i$ for each $i \in \mathcal{P}_n$. Let $\overline{t_n}(\pi)$ be the number of experiments counted in $t_n(\pi)$, where $\omega_n = 1$ also holds. We then compute $\check{\alpha}'_n(\overline{\pi})$ using the equation

$$\check{\alpha}'_n(\overline{\pi}) = \begin{cases} \overline{t_n}(\pi)/t_n(\pi) & \text{if } \overline{\pi}_n = 1, \\ 1 - \overline{t_n}(\pi)/t_n(\pi) & \text{if } \overline{\pi}_n = 0. \end{cases} \tag{3}$$

The vector $\overline{\pi} \in \{0,1\}^{\overline{\mathcal{P}_n}}$ is added to G_n if

$$\check{\alpha}'_n(\overline{\pi}) \hat{\beta}_n(\pi, \hat{A}_{n,\pi}) \leq 2eS(\lambda), \tag{4}$$

where $S(\lambda)$ is defined as

$$S(\lambda) := \frac{12\lambda N^2 C \log T}{T}.$$

This G_n reserves such $\overline{\pi} \in \{0,1\}^{\overline{\mathcal{P}_n}}$ that $\check{\alpha}'_n(\overline{\pi})$ is too small

to estimate $\alpha_n(\overline{\pi})$ with sufficient accuracy. Then $\check{\alpha}_n(\overline{\pi})$ is determined by replacing $\check{\alpha}'_n(\overline{\pi})$ with 0 for $\overline{\pi} \in G_n$:

$$\check{\alpha}_n(\overline{\pi}) := \begin{cases} \check{\alpha}'_n(\overline{\pi}) & \text{if } \overline{\pi} \notin G_n, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

This replacement contributes to reducing the relative estimation error of $\hat{\beta}_{n'}$ in subsequent steps ($n' = n+1, \dots, N$).

After iterating for all $n = 1, 2, \dots, N$, the algorithm computes H_n and D_n ($n = 1, 2, \dots, N$) defined by

$$H_n = \left\{ \overline{\pi} \in \{0,1\}^{\overline{\mathcal{P}_n}} \,\middle|\, \hat{\beta}_n(\overline{\pi}_{\mathcal{P}_n}, \hat{A}_{n,\pi}) \le 8eC^2 S(\lambda) \right\}, \tag{6}$$

$$D_n = G_n \cup H_n. \tag{7}$$

This D_n contributes to bound the absolute error of the estimation of $\hat{\beta}_n(\overline{\pi}_{\mathcal{P}_n})$ for $\overline{\pi} \notin D_n$. The algorithm returns an estimate $\hat{\beta}_n$ and the family $D := \{D_n \mid n = 1, 2, \dots, N\}$.

3.3. Second Phase: Estimation of α

In this phase, our algorithm computes an estimate $\hat{\alpha}_n$ of α_n for all $n = 1, \dots, N$. The pseudo-code for this phase is given in Algorithm 2. As an input, it receives $\hat{\beta}_n$ ($n = 1, \dots, N$) and D from Algorithm 1.

Algorithm 2 consists of two parts. The first part conducts $T/(3C)$ experiments with $\hat{A}_{n,\pi}$ (computed from $\hat{\beta}_n(\pi, A)$, $A \in \mathcal{A}$) for each $n = 1, \dots, N$ and $\pi \in \{0,1\}^{\mathcal{P}_n}$. This is the same process used to compute $\check{\alpha}'_n$ in Algorithm 1. Let

$$D_n^{\downarrow} := \{\pi \in \{0,1\}^{\mathcal{P}_n} \mid \overline{\pi}^0, \overline{\pi}^1 \in D_n\}$$

where $\overline{\pi}^k$ is the extension of $\pi \in \{0,1\}^{\mathcal{P}_n}$ onto $\{0,1\}^{\overline{\mathcal{P}_n}}$ with $\overline{\pi}^k_n = k$. Let us define a set $J_n := \{0,1\}^{\mathcal{P}_n} \setminus D_n^{\downarrow}$ and a constant $r_{n,\pi} := \hat{\beta}_n(\pi, \hat{A}_{n,\pi})/C$ for each $n = 1, \dots, N$ and $\pi \in \{0,1\}^{\mathcal{P}_n}$. In the second part, the algorithm solves the following optimization problem:

$$\min_{\eta \in [0,1]^{\mathcal{A}}} \max_{A \in \mathcal{A}} \sum_{n \in I_{N,A}} \sum_{\pi \in J_n} \frac{\hat{\beta}_n^2(\pi, A)}{\sum_{A' \in \mathcal{A}} \eta_{A'} \hat{\beta}_n(\pi, A') + r_{n,\pi}}$$
$$\text{s.t.} \sum_{A' \in \mathcal{A}} \eta_{A'} = 1. \tag{8}$$

Note that, for each $n = 1, 2, \dots, N$, $\pi \in J_n$ only if $\hat{\beta}_n(\pi, \hat{A}_{n,\pi}) > 0$ according to Line 20 of Algorithm 1. Thus the denominator is positive for every $\pi \in J_n$, and the above optimization problem is well-defined. Let $\hat{\eta}$ be an optimal solution for (8). Consider the distribution $\mathcal{U}(\hat{\eta})$ over $\mathcal{A}$ that generates A with a probability of $\hat{\eta}_A$. The second part samples an intervention according to $\mathcal{U}(\hat{\eta})$ and uses it to conduct experiments, for $T/3$ times,.

For each $n = 1, \dots, N$ and $\pi \in \{0,1\}^{\mathcal{P}_n}$, the algorithm counts the number $t'_n(\pi)$ (resp., $\overline{t'_n}(\pi)$) of experiments that result in $\omega \in \{0,1\}^N$ with $\omega_{\mathcal{P}_n} = \pi$ (resp., $\omega_{\mathcal{P}_n} = \pi$ and $\omega_n = 1$). Then, $\hat{\alpha}'_n(\pi)$ ($n = 1, \dots, N$, $\pi \in \{0,1\}^{\overline{\mathcal{P}_n}}$) is defined by

$$\hat{\alpha}'_n(\pi) = \begin{cases} \overline{t'_n}(\pi_{\mathcal{P}_n})/t'_n(\pi_{\mathcal{P}_n}) & \text{if } \pi_n = 1, \\ 1 - \overline{t'_n}(\pi_{\mathcal{P}_n})/t'_n(\pi_{\mathcal{P}_n}) & \text{if } \pi_n = 0. \end{cases} \tag{9}$$

The output $\hat{\alpha}_n$ is defined by

$$\hat{\alpha}_n(\pi) = \begin{cases} \hat{\alpha}'_n(\pi) & \text{if } \pi \notin D_n, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

3.4. Regret bound

Pseudo-code of our entire algorithm is provided in Algorithm 3. It computes an estimate $\hat{\beta}$ of β by Algorithm 1 and then computes $\hat{\alpha}$ by Algorithm 2. It then computes an estimate $\hat{\mu}$ of μ by

$$\hat{\mu}(A) = \sum_{\pi \in B(A)} \prod_{n \in I_{N,A}} \hat{\alpha}_n(\pi_{\overline{\mathcal{P}_n}}) \tag{11}$$

for each $A \in \mathcal{A}$. The algorithm returns an intervention $\hat{A} \in \mathcal{A}$ that maximizes $\hat{\mu}$.

Let us define γ^* as the optimum value of the following problem:

$$\gamma^* := \min_{\eta \in [0,1]^{\mathcal{A}}} \max_{A \in \mathcal{A}} \sum_{n=1}^{N} \sum_{\substack{\pi \in \{0,1\}^{\mathcal{P}_n} \\ : \beta_n(\pi, A) > 0}} \frac{\beta_n^2(\pi, A)}{\sum_{A' \in \mathcal{A}} \eta_{A'} \beta_n(\pi, A')}$$
$$\text{s.t.} \sum_{A' \in \mathcal{A}} \eta_{A'} = 1. \tag{12}$$

The regret bound of Algorithm 3 is parameterized by the optimum value γ^*:

Theorem 1. *The regret R_T of Algorithm 3 satisfies*

$$R_T \le O\left(\sqrt{\frac{\max\{\gamma^*, N\} \log(|\mathcal{A}|T)}{T}}\right).$$

The notation $O(\cdot)$ is used here under the assumption that N is sufficiently small with respect to T but not negligible. The optimum value γ^* is bounded as follows. Let $|A|$ denote the number of nodes intervened by A, i.e., $|A| := |\{n \in [1, N] \colon A_n \in \{0,1\}\}|$:

Proposition 2. *It holds that* $N - \min_{A \in \mathcal{A}} |A| \le \gamma^* \le \min\{NC, N|\mathcal{A}|\}$.

Since the lower-bound for the general best-arm identification problem is $\Omega(\sqrt{|\mathcal{A}|/T})$ (Audibert & Bubeck, 2010)[Theorem 4], our algorithm provides a better regret bound when the number of interventions $|\mathcal{A}|$ is large compared to $\gamma^* \le NC$, which is only dependent on the causal graph structure.

Remark 3. We present Algorithms 1, 2, and 3 for the setting that every $\alpha_n(\pi)$ is unknown. However, our algorithms can be applied even when $\alpha_n(\pi)$ is known for some $n = 1, \dots, N$ and $\pi \in \{0,1\}^{\overline{\mathcal{P}_n}}$ by incorporating minor modifications. In this case, we denote the number of unknown $\alpha_n(\pi)$ as C. The modified algorithm just skips experiments for estimating the known $\alpha_n(\pi)$, and we can define $\hat{\beta}_n(\pi_{\mathcal{P}_n}, A) = 0$ for such n and π. We then redefine γ^* by replacing corresponding $\beta_n(\pi_{\mathcal{P}_n}, A)$ with 0 in (12), and our bound in Theorem 1 is valid for this decreased γ^*. In particular, we can recover the regret bound considered in (Lattimore et al., 2016)[Theorem 3] as follows:

Corollary 4. *Suppose that $\alpha_n(\pi)$ is known for every $n < N$ and $\pi \in \{0,1\}^{\overline{\mathcal{P}_n}}$. Then the regret R_T of Algorithm 3 satisfies $R_T \leq O(\sqrt{\gamma^* \log(|\mathcal{A}|T)/T})$, where*

$$\gamma^* = \min_{\eta \in [0,1]^{\mathcal{A}}} \max_{A \in \mathcal{A}} \sum_{\pi \in \{0,1\}^{\mathcal{P}_N}} \frac{\beta_N^2(\pi, A)}{\sum_{A' \in \mathcal{A}} \eta_{A'} \beta_N(\pi, A')}$$
$$\text{s.t.} \sum_{A' \in \mathcal{A}} \eta_{A'} = 1.$$

Remark 5. Our problem setting is often called *hard intervention*, which directly controls the realization of a node v_n as $A_n \in \{0,1\}$. In contrast, Sen et al. (2017) introduced the *soft intervention* model on a node v_n where an intervention changes the conditional probability α_n of a node v_n. They in fact considered a simple case where a graph has a single node v_k such that $\mathcal{P}_N = \mathcal{P}_k \cup \{k\}$, whose conditional probability can be controlled by soft intervention, and proved parameterized regret bound.

We here remark that their model can be implemented by the hard intervention model with an arbitrary set of interventions. The details of this implementation is presented in our full paper.

4. Proofs

This section presents an approach for proving Theorem 1. Complete proofs for all the statements are presented in the full version (Yabe et al., 2018).

4.1. Accuracy of Algorithm 1

For $n = 1, 2, \dots, N$, let $\check{\alpha}_n$ and $\check{\alpha}'_n$ be the stochastic estimates computed in Algorithm 1, and $\hat{A}_{n,\pi} := \operatorname{argmax}_{A \in \mathcal{A}} \hat{\beta}_n(\pi, A)$ be the action determined from the estimate $\hat{\beta}_n$. Let G be defined by $G = \{G_n \mid n = 1, \dots, N\}$. Using G, we define $\alpha_{n,G}$ and $\beta_{n,G}$ as follows. For each $n \in [1, N]$ and $\overline{\pi} \in \{0,1\}^{\overline{\mathcal{P}_n}}$, we define $\alpha_{n,G}(\overline{\pi})$ by

$$\alpha_{n,G}(\overline{\pi}) := \begin{cases} \alpha_n(\overline{\pi}) & \text{if } \overline{\pi} \notin G_n, \\ 0 & \text{otherwise.} \end{cases}$$

For each $n \in [1, N]$, $\pi \in \{0,1\}^{\mathcal{P}_n}$, and $A \in \mathcal{A}$, we define $\beta_{n,G}(\pi, A)$ by

$$\beta_{n,G}(\pi, A) = \sum_{\pi' \in B_n(\pi, A)} \prod_{m \in I_{n-1,A}} \alpha_{m,G}(\pi'_{\overline{\mathcal{P}_n}}).$$

Thus $\alpha_{n,G}(\overline{\pi})$ is obtained from $\alpha_n(\overline{\pi})$ by truncating its values if $\overline{\pi} \in G_n$, and $\beta_{n,G}$ is defined from $\alpha_{n,G}$. We define $\alpha_{n,D}$ and $\beta_{n,D}$ in the same way. Since $G_n \subseteq D_n$, we observe that $\beta_{n,D}(\pi, A) \leq \beta_{n,G}(\pi, A) \leq \beta_n(\pi, A)$. Similarly, for $A \in \mathcal{A}$, we define $\mu_D(A)$ by

$$\mu_D(A) = \sum_{\pi \in B(A)} \prod_{m \in I_{N,A}} \alpha_{m,D}(\pi_{\overline{\mathcal{P}_m}}). \tag{13}$$

The following proposition demonstrates the error bound for outputs $\hat{\beta}_n$ and D from Algorithm 1.

Proposition 6. *Let $\hat{\beta}_n$ and D be the outputs of Algorithm 1 with parameter $\lambda \geq 1$. Then the following holds with a probability of at least $1 - 6C/T$: for every $n \in [1, N]$, $\overline{\pi} \in \{0,1\}^{\overline{\mathcal{P}_n}} \setminus D_n$ with $\pi = \overline{\pi}_{\mathcal{P}_n}$, and $A \in \mathcal{A}$:*

$$\frac{1}{e}\beta_{n,D}(\pi, A) \leq \hat{\beta}_n(\pi, A) \leq e\beta_n(\pi, A), \tag{14}$$
$$\alpha_n(\overline{\pi})\beta_n(\pi, \hat{A}_{n,\pi}) \geq S(\lambda), \tag{15}$$
$$\beta_n(\pi, A) \leq e\hat{\beta}_n(\pi, A) + e\hat{\beta}_n(\pi, \hat{A}_{n,\pi})/C, \tag{16}$$
$$\mu(A) - \mu_D(A) \leq 8e^2(C^3 + C)S(\lambda). \tag{17}$$

We prepare the following three lemmas to prove Proposition 6. The first lemma is an application of Chernoff's bound, which bounds the relative error of the estimation $\check{\alpha}'_n$:

Lemma 7. *Let $n \in [1, N]$, $\overline{\pi} \in \{0,1\}^{\overline{\mathcal{P}_n}}$, and $\pi = \overline{\pi}_{\mathcal{P}_n}$.*

(i) If $\alpha_n(\overline{\pi})\beta_n(\pi, \hat{A}_{n,\pi}) \leq S(\lambda)$, then the following holds with a probability of at least $1 - 2/T$:

$$\check{\alpha}'_n(\overline{\pi})\beta_n(\pi, \hat{A}_{n,\pi}) \leq 2S(\lambda).$$

(ii) If $\alpha_n(\overline{\pi})\beta_n(\pi, \hat{A}_{n,\pi}) \geq S(\lambda)$, then the following holds with a probability of at least $1 - 3/T$:

$$\left(1 - \frac{1}{\sqrt{\lambda}N}\right)\alpha_n(\overline{\pi}) \leq \check{\alpha}'_n(\overline{\pi}) \leq \left(1 + \frac{1}{\sqrt{\lambda}N}\right)\alpha_n(\overline{\pi}).$$

The second lemma bounds the gap produced by truncation of α_n that is conducted for introducing $\alpha_{n,G}$ and $\alpha_{n,D}$. We use the notation $H_n^{\downarrow} := \{\pi_{\mathcal{P}_n} \mid \pi \in H_n\}$.

Lemma 8. *(i) Let $n \in [1, N]$ and $\pi \in \{0,1\}^{\mathcal{P}_n}$. For every $A \in \mathcal{A}$, it holds that*

$$\beta_n(\pi, A) - \beta_{n,G}(\pi, A)$$
$$\leq \sum_{m=1}^{N} \sum_{\pi' \in G_m} \max_{A' \in \mathcal{A}} \alpha_m(\pi')\beta_{m,G}(\pi'_{\mathcal{P}_m}, A').$$

(ii) For every $A \in \mathcal{A}$, it holds that

$$\mu(A) - \mu_D(A) \leq \sum_{m=1}^{N} \sum_{\pi \in G_m} \max_{A' \in \mathcal{A}} \alpha_m(\pi) \beta_{m,G}(\pi_{\mathcal{P}_m}, A') + \sum_{m=1}^{N} \sum_{\pi' \in H_m^{\downarrow}} \max_{A'' \in \mathcal{A}} \beta_{m,G}(\pi', A'').$$

The third lemma bounds the relative error of $\hat{\beta}$. This statement can be proven by induction on the basis of Lemma 7.

Lemma 9. *The following holds for every $n = 1, 2, \ldots, N$, $\overline{\pi} \in \{0,1\}^{\overline{\mathcal{P}_n}}$ with $\pi = \overline{\pi}_{\mathcal{P}_n}$, and $A \in \mathcal{A}$ with a probability of at least $1 - 6C/T$:*

$$\frac{1}{e}\beta_{n,G}(\pi, A) \leq \hat{\beta}_n(\pi, A) \leq e\beta_{n,G}(\pi, A),$$
$$\overline{\pi} \in G_n \quad \text{if } \alpha_n(\overline{\pi})\beta_n(\pi, \hat{A}_{n,\pi}) < S(\lambda),$$
$$\left(1 - \frac{1}{\sqrt{\lambda}N}\right)\alpha_n(\overline{\pi}) \leq \check{\alpha}'_n(\overline{\pi}) \leq \left(1 + \frac{1}{\sqrt{\lambda}N}\right)\alpha_n(\overline{\pi}) \quad \text{if } \alpha_n(\overline{\pi})\beta_n(\pi, \hat{A}_{n,\pi}) \geq S(\lambda).$$

Then Proposition 6 is proven on the basis of Lemmas 7–9.

4.2. Accuracy of Algorithm 2

This subsection bounds the gap between the true value $\mu(A)$ and its estimate $\hat{\mu}(A)$ given by Algorithm 2, assuming that the input of Algorithm 2, which is output of Algorithm 1, satisfies the conditions in Proposition 6.

Proposition 10. *Suppose that $\lambda \geq 1$, and $\hat{\beta}_n$ and D satisfy (14), (15), (16), and (17). Let $\hat{\alpha}_n$ be the output of Algorithm 2, and let $\hat{\mu}$ be defined by (11). Then the following holds for every $A \in \mathcal{A}$ with a probability of at least $1 - (10C + 2)/T$:*

$$|\mu(A) - \hat{\mu}(A)| \leq \sqrt{\frac{2e^6\gamma^* \log(|\mathcal{A}|T)}{T}} + \sqrt{\frac{8e^2C^3 \log T}{\lambda T}} + 8e^2(C^3 + C)S(\lambda).$$

Recall that $I_{N,A} := \{m \in [1, N] \mid A_m = *\}$ for $A \in \mathcal{A}$. For $n \in [1, N]$ and $\pi \in \{0,1\}^{\overline{\mathcal{P}}_n}$, let $\Delta\alpha_n(\pi) := \hat{\alpha}_n(\pi) - \alpha_{n,D}(\pi)$. For $A \in \mathcal{A}$ and $J \subseteq I_{N,A}$, we define $f^J(A)$ by

$$f^J(A) = \sum_{\pi \in B(A)} \prod_{m \in I_{N,A} \setminus J} \alpha_{m,D}(\pi_{\overline{\mathcal{P}_m}}) \prod_{n \in J} \Delta\alpha_n(\pi_{\overline{\mathcal{P}_n}}).$$

Observe that $f^J(A)$ is given by replacing $\alpha_{n,D}(\pi_{\overline{\mathcal{P}_n}})$ by $\Delta\alpha_n(\pi_{\overline{\mathcal{P}_n}})$ for $n \in J$ in the definition (13) of μ_D. Recall that $\hat{\mu}(A)$ is given by replacing $\alpha_{n,D}(\pi)$ in the definition of μ_D by $\hat{\alpha}_n(\pi)$ for all $n \in I_{N,A}$. Based on these relationships, we have the following lemma:

Lemma 11. *For $A \in \mathcal{A}$, it holds that:*

$$\mu_D(A) = f^{\emptyset}(A), \quad \hat{\mu}(A) = \sum_{J \subseteq I_{N,A}} f^J(A). \tag{18}$$

For $j \in I_{N,A}$, let $f^j(A) := f^{\{j\}}(A)$. We provide probabilistic bounds for the linear terms ($|J| = 1$) and super-linear terms ($|J| \geq 2$) in (18), separately, using Hoeffding's inequality.

Lemma 12. *Suppose that (14), (15), and (16) hold.*

(i) The following holds with a probability of at least $1 - (C + 2)/T$:

$$\max_{A \in \mathcal{A}} \left| \sum_{j \in I_{N,A}} f^j(A) \right| \leq \sqrt{\frac{2e^6\gamma^* \log(|\mathcal{A}|T)}{T}}.$$

(ii) The following holds with a probability of at least $1 - 9C/T$:

$$\max_{A \in \mathcal{A}} \sum_{J \subseteq I_{N,A} : |J| \geq 2} |f^J(A)| \leq \sqrt{\frac{8e^2C^3 \log T}{\lambda T}}.$$

The above two lemmas imply Proposition 10.

4.3. Proof of Theorem 1

We present a sketch of proof of Theorem 1, on the basis of Propositions 6 and 10, as follows. Putting $\lambda = C^3/N$, by Propositions 2 6, and 10, the following holds for every $A \in \mathcal{A}$ with a probability of at least $1 - (16C + 2)/T$:

$$|\mu(A) - \hat{\mu}(A)| \leq \sqrt{\frac{8e^6 \max\{\gamma^*, N\} \log(|\mathcal{A}|T)}{T}} + \frac{192e^2NC^7 \log T}{T}.$$

Let $A^* = \mathrm{argmax}_{A \in \mathcal{A}} \mu(A)$ and $\hat{A} = \mathrm{argmax}_{A \in \mathcal{A}} \hat{\mu}(A)$. Then it holds that

$$\mu(A^*) - \mu(\hat{A}) \leq |\mu(A^*) - \hat{\mu}(A^*)| + |\mu(\hat{A}) - \hat{\mu}(\hat{A})| = O\left(\sqrt{\frac{\max\{\gamma^*, N\} \log(|\mathcal{A}|T)}{T}}\right).$$

This implies the desired regret bound.

5. Experiments

We now demonstrate the performance of the proposed algorithm through experimental evaluations and compare it with a baseline algorithm (Audibert & Bubeck, 2010) which was proposed for the general bandit problem and thus cannot take advantage of known causal graph structure.

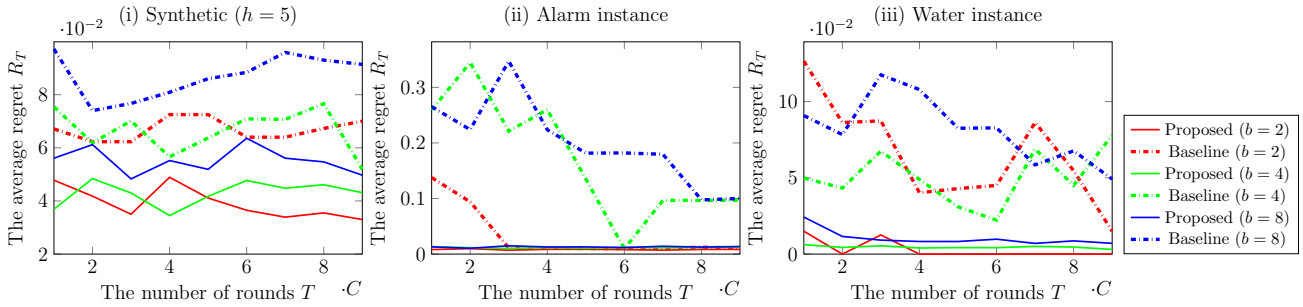

Figure 2. The average regret over synthetic and real-world instances

Instances We evaluated the algorithms on both synthetic and real-world instances. Detailed experimental setting is presented in the full version (Yabe et al., 2018). Recall that an instance of the causal bandit problem consists of a DAG G, an intervention set $\mathcal{A}$, and $\alpha_n (n = 1, \dots, N)$.

In the synthetic instances, the DAG G is defined as a directed complete binary tree of height 4, and then the number of nodes is $N = 2^5 - 1 = 31$, and the number of uncertain parameter is $C = 2^2 \times (2^4 - 1) + 2^0 \times 2^4 = 76$. In the real-world instances, the DAG G is constructed from the Alarm and the Water data sets in a Bayesian Network Repository[1]. The numbers N of nodes in the DAGs constructed from Alarm and Water data sets are 37 and 32, and the numbers C of uncertain parameters are 116 and 248, respectively.

For each G, we consider interventions over all leaves which fixes exactly $b \in \mathbb{N}$ nodes as 1 and the others as 0. We call this parameter b budget, and the number of intervention $|\mathcal{A}|$ is then controlled by the budget.

For each $n \in \{1, \dots, N\}$ and $\pi \in \{0, 1\}^{\overline{\mathcal{P}_n}}$, we generate $\alpha_n(\pi)$ from the uniform distribution over $[0, 1]$.

For each of those instances, we executed the algorithms 10 times and compared their average regrets.

Implementation of the proposed algorithm Our algorithm given in Section 3 is designed conservatively to obtain the theoretical regred bound (Theorem 1), and there is a room to modify the algorithm to be more efficient in practice although the theoretical regret bound may not hold for it. In our implementation, we introduced the following three modifications into the proposed algorithm. First, while Algorithm 2 discards samples obtained for computing $\check{\alpha}'$ in Algorithm 1 to maintain the independence between $\hat{\beta}$ and $\hat{\alpha}$, we use all of them also in Algorithm 2 in our implementation. Next, we ignore the truncation mechanism of Algorithm 1 by setting $\lambda = 0$. We expect these two modifications make the estimates of the algorithm more accurate. Finally, instead of solving (8), we set η_A by $\eta_A = 1/C$ if $A = \hat{A}_{n,\pi}$ for some $n \in [1, N]$ and $\pi \in \{0, 1\}^{\mathcal{P}_n}$, and $\eta_A = 0$ otherwise. Since it is time-consuming to solve (8), this modification makes the algorithm faster.

[1] http://www.cs.huji.ac.il/~galel/Repository/

Experimental results Figure 4.3(i) shows the average regrets over the synthetic instances against the number of rounds $T \in \{C, 2C, \dots, 9C\}$. Figures 4.3 (ii) and (iii) respectively illustrate the average regrets for the real-world instances constructed from the Alarm and the Water data sets.

The results show that the proposed algorithm outperforms the baseline in every instance. In particular, the gap is remarkably large (> 0.2) in the Alarm data set (ii) with a large number of interventions ($b = 4, 8$, corresponding to $|\mathcal{A}| = 793, 3796$, respectively,) and a small number of samples ($T \leq 4C = 464$). In these cases, the baseline cannot apply every intervention at least once. On the other hand, the regret of the proposed algorithm only grows slowly with respect to the number of arms $|\mathcal{A}|$, in all instances. Thus the proposed algorithm provides effective regret, even when the number of interventions $|\mathcal{A}|$ is 30 times larger than the number of experiments T.

6. Conclusion

In this paper, we proposed the first algorithm for the general causal bandit problem, where existing algorithms could deal with only localized interventions, and proved a novel regret bound $O(\sqrt{\gamma^* \log(|\mathcal{A}|T)/T})$ which is logarithmic with respect to the number of arms. Our experimental result shows that the proposed algorithm is applicable to systems where the number of interventions $|\mathcal{A}|$ is much larger than T. One important future research direction would be to prove the gap-dependent bound as Sen et al. (2017) has proven for localized interventions. Another research direction, which is mentioned in (Lattimore et al., 2016), would include incorporation of a causal discovery algorithm to enable the estimation of the structure of a causal graph, which is currently assumed to be known in advance.

Acknowledgements

Daisuke Hatano, Hanna Sumita, Naonori Kakimura, Takuro Fukunaga, and Ken-ichi Kawarabayashi are supported by JST ERATO Kawarabayashi Large Graph Project, Grant Number JPMJER1201, Japan.

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646, 2014.

Audibert, J.-Y. and Bubeck, S. Best arm identification in multi-armed bandits. In *The 23rd Conference on Learning Theory*, pp. 41–53, 2010.

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

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

Eberhardt, F., Glymour, C., and Scheines, R. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pp. 178–184. AUAI Press, 2005.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pp. 3212–3220, 2012.

Hauser, A. and Bühlmann, P. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4): 926–939, 2014.

Hu, H., Li, Z., and Vetta, A. R. Randomized experimental design for causal graph discovery. In *Advances in Neural Information Processing Systems*, pp. 2339–2347, 2014.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17 (1):1–42, 2016.

Kim, D. J., Ferrin, D. L., and Rao, H. R. A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision Support Systems*, 44(2):544–564, 2008.

Lattimore, F., Lattimore, T., and Reid, M. D. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*, pp. 1181–1189, 2016.

Meinshausen, N., Hauser, A., Mooij, J. M., Peters, J., Versteeg, P., and Bühlmann, P. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27): 7361–7368, 2016.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Pearl, J. *Causality*. Cambridge university press, 2009.

Sen, R., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. Identifying best interventions through online importance sampling. In *International Conference on Machine Learning*, pp. 3057–3066, 2017.

Shanmugam, K., Kocaoglu, M., Dimakis, A. G., and Vishwanath, S. Learning causal graphs with small interventions. In *Advances in Neural Information Processing Systems*, pp. 3195–3203, 2015.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12(Apr):1225–1248, 2011.

Spirtes, P. and Glymour, C. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pp. 465–472, 1990.

Yabe, A., Hatano, D., Sumita, H., Ito, S., Kakimura, N., Fukunaga, T., and Kawarabayashi, K. Causal bandits with propagating inference. *arXiv preprint arXiv:1806.02252*, 2018.