# Information-Directed Sampling for Causal Bandits

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Causal bandits exploit structural relationships among variables to share information across interventions and accelerate the identification of high-reward decisions. In many applications, however, some variables cannot be directly manipulated, even though they influence the reward and provide useful information about the underlying causal system. We study contextual causal bandits with non-manipulable variables, where context variables are observed before action selection and additional variables are observed after each intervention. Assuming a known causal graph without latent confounding, we adopt a Bayesian formulation in which the conditional probability tables of the observational distribution constitute the unknown parameter. This representation allows observations collected under one intervention to update reward estimates for other interventions through their shared causal mechanisms. We develop causal variants of Thompson Sampling and Information-Directed Sampling (IDS) for this setting. For Thompson Sampling, we establish an entropy-dependent sublinear Bayesian regret bound. For IDS, we derive an entropy-dependent regret bound that explicitly quantifies the additional error introduced by Monte Carlo approximation of the expected regret and information gain; when these quantities are available exactly, the bound recovers the standard sublinear IDS rate. We further provide high-probability confidence bounds for the Monte Carlo estimates used by the algorithm. Experiments on several synthetic causal bandit tasks show that the proposed methods outperform causal and non-causal baselines by more effectively exploiting information shared across interventions.

## 1 Introduction

In the classical multi-armed bandit problem, a decision-maker repeatedly selects an action from a finite set and observes the resulting reward. Because the reward distributions are initially unknown, the learner must balance *exploration*, which gathers information about uncertain actions, and *exploitation*, which favors actions believed to yield high rewards Lattimore & Szepesvári (2020); Slivkins et al. (2019); Vermorel & Mohri (2005). Without additional structure, observing the reward of one action typically provides no direct information about the rewards of the other actions. Consequently, effective algorithms must explore the available actions sufficiently often Garivier & Cappé (2011); Jamieson & Nowak (2014); Jamieson et al. (2014).

In many decision-making problems, however, actions are related through a common underlying mechanism. An observation collected after one action may therefore provide information about several other actions. Structured bandit methods exploit such relationships to improve statistical efficiency and reduce unnecessary exploration Schulz et al. (2020); Jun & Zhang (2020); Tirinzoni et al. (2020); Van Parys & Golrezaei (2024); Wan et al. (2023); Mersereau et al. (2009). The central challenge is to characterize and exploit the relevant information-sharing structure while continuing to balance exploration and exploitation.

Causal bandits provide a principled framework for structured decision-making when the environment is governed by an underlying structural causal model Lattimore et al. (2016); Sen et al. (2017); Lee & Bareinboim (2018); Wei et al. (2024); Qasim Elahi et al. (2024). In this setting, each action corresponds to an intervention on the causal system, and the resulting observations are generated according to a shared causal model Pearl (2009). Unlike an unstructured bandit model, the reward distributions associated with different interventions are coupled through common causal mechanisms. Consequently, observations obtained under one intervention

can potentially improve the learner's estimates of the rewards associated with other interventions Lattimore et al. (2016); Sen et al. (2017); Yabe et al. (2018); Lee & Bareinboim (2018).

Many existing causal bandit formulations permit interventions on a large collection of observed variables. In practice, however, some variables cannot be directly manipulated. Examples include genetic characteristics in healthcare, demographic attributes in public-policy applications, and macroeconomic conditions in economic decision-making. Although such variables are non-manipulable, they may strongly influence the reward and the effects of feasible interventions. Their presence therefore changes the set of candidate interventions and creates an additional challenge for efficiently sharing information across actions.

Lee and Bareinboim Lee & Bareinboim (2019) study causal bandits with non-manipulable variables in graphs that may contain latent confounding. They characterize the interventions that can be optimal under the manipulability constraints and use a generalized $z^2$-identification procedure to derive multiple estimators of interventional reward distributions. These estimators are combined through a bootstrap-based minimum-variance weighted average and incorporated into variants of Thompson Sampling and KL-UCB. Their results demonstrate the empirical value of exploiting causal information, but do not provide regret guarantees for the resulting algorithms.

We study contextual causal bandits with non-manipulable variables under the assumption that the causal graph is known and contains no latent confounders. At each round, context variables are observed before the learner selects an intervention. The learner subsequently observes the reward and the remaining observed variables in the causal graph. For example, in a healthcare application, patient characteristics and medical history may be available before treatment selection, whereas physiological measurements that cannot be directly manipulated may be observed after treatment. The learner's objective is to select a feasible intervention that maximizes the context-dependent expected reward.

Following the Bayesian information-theoretic framework of Russo & Van Roy (2016; 2014), we represent uncertainty about the causal system using a random parameter $\boldsymbol{\theta}$. Because the graph contains no latent confounders, the relevant interventional distributions are identifiable from the observational distribution through the truncated factorization formula. We therefore let $\boldsymbol{\theta}$ collect the conditional probability tables associated with the causal graph. Observations from each round update the posterior distribution over these shared parameters, allowing samples collected under one intervention to improve the estimated rewards and information gains of other interventions.

Based on this formulation, we develop causal variants of Thompson Sampling and Information-Directed Sampling. Thompson Sampling selects an intervention according to its posterior probability of being optimal under the observed context. IDS instead selects a distribution over interventions by balancing their expected instantaneous regret against the information they provide about the context-dependent optimal decision. Since the posterior expectations required by IDS are generally unavailable in closed form, we estimate them using Monte Carlo samples and explicitly account for the resulting approximation error.

Our main contributions are summarized as follows:

- We formulate contextual causal bandits with non-manipulable variables in a Bayesian framework in which the conditional probability tables of the observational distribution are treated as the unknown parameter. This formulation enables observations collected under one intervention to update estimates associated with other interventions through their shared causal mechanisms.

- We propose causal Thompson Sampling and Information-Directed Sampling algorithms for this setting. We establish an entropy-dependent sublinear Bayesian regret bound for Thompson Sampling. For IDS, we derive a regret bound that separates the standard information-theoretic term from the additional error caused by Monte Carlo approximation; the oracle version, in which the information ratio is computed exactly, achieves the standard sublinear IDS guarantee.

- We derive high-probability concentration bounds for the Monte Carlo estimates of the expected instantaneous regret and contextual information gain. These results provide computable confidence sets and quantify the effect of posterior-sampling error on the IDS regret guarantee.

- We evaluate the proposed methods on multiple synthetic causal bandit tasks, including structured examples and randomly generated causal graphs. The results show that the proposed algorithms outperform causal and non-causal baselines by more effectively exploiting information shared across interventions.

## 2 Preliminaries

We adopt the Structural Causal Model (SCM) framework Pearl (2009). An SCM, denoted by $\mathcal{M}$, is defined as a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$, where $\mathbf{U}$ denotes a set of exogenous (unobserved) variables determined by factors external to the model, and $\mathbf{V}$ denotes a set of endogenous (observed) variables determined by variables in $\mathbf{U} \cup \mathbf{V}$ through the structural functions $\mathbf{F}$. In our setting, the endogenous variables $\mathbf{V}$ take values in finite domains and consist of the reward variable $Y$, actionable (manipulable) variables, non-actionable (non-manipulable) variables, and context variables. For further simplicity, we assume they are binary throughout the paper. For any subset of nodes $\mathbf{X} \subseteq \mathbf{V}$, let $\Omega(\mathbf{X})$ denote the Cartesian product of the state spaces of all variables in $\mathbf{X}$. However, our proposed methods and results also hold when the observed nodes are discrete and may take more than two possible values. The structural functions $\mathbf{F}$ specify how each $V_i$ is assigned a value, denoted as $v_i = f_i(\mathbf{PA}^i, \mathbf{U}^i)$, based on the values of its parent variables $\mathbf{PA}^i \subseteq \mathbf{V}$ and exogenous variables $\mathbf{U}^i \subseteq \mathbf{U}$. Finally, $P(\mathbf{U})$ is the probability distribution over the exogenous variables $\mathbf{U}$.

Each SCM is associated with a causal graph $\mathcal{G} = \langle \mathbf{V}, \mathbf{E} \rangle$, where the edge set $\mathbf{E}$ consists of two types: directed edges, such as $V_i \to V_j$, which indicate direct functional dependence (i.e., $V_i$ is used in defining $f_j$ in $\mathbf{F}$), and bidirected edges, such as $V_i \leftrightarrow V_j$, which represent the presence of an unobserved (latent) confounder affecting both $V_i$ and $V_j$. We use notations *pa*, *ch*, *an*, and *de* to refer to the parents, children, ancestors, and descendants of a variable, respectively. Capitalized forms, such as *Pa*, *Ch*, *An*, and *De*, include the variable itself (e.g., $An(W) = an(W) \cup W$). For a set of variables, the relations are defined as the union of their individual outputs, e.g., $An(\mathbf{W}) = \bigcup_{W \in \mathbf{W}} An(W)$. Note that $pa(V_i) = \mathbf{PA}^i$. A subgraph of $\mathcal{G}$, denoted $\mathcal{G}_{\overline{\mathbf{X}}}$, is obtained by removing edges pointing to the variables in $\mathbf{X}$. The connected component (c-component) of the DAG $\mathcal{G}$, containing vertex $V_i$, is denoted by $\mathsf{CC}_{\mathcal{G}}(V_i)$, which is the maximal set of all vertices in $\mathcal{G}$ that have a path to $V_i$, consisting only of bi-directed edges Tian & Pearl (2002).

In the $K$-armed bandit problem, $K$ arms with distinct reward distributions are available, and the goal is to minimize cumulative regret over $T$ rounds. Regret is defined as the difference between the maximum expected cumulative reward achievable by always selecting the optimal arm and the expected cumulative reward obtained by a given algorithm. In the SCM-MAB setting, each arm corresponds to an intervention on a subset of variables. Given a causal graph $\mathcal{G}$ with reward $Y$, the arms are defined as $\text{do}(\mathbf{X} = \mathbf{x}) \mid \mathbf{X} \subseteq \mathbf{V} \setminus Y$, where the distribution of the reward variable under the intervention $\text{do}(\mathbf{X} = \mathbf{x})$, denoted by $P(Y_{\mathbf{x}})$, coincides with the interventional distribution $P_{\mathbf{x}}(Y)$. The expected reward associated with an intervention is $\mu_{\mathbf{x}} = \mathbb{E}[Y \mid \text{do}(\mathbf{x})]$. When additional context variables $\mathbf{C}$ are observed prior to the intervention, the objective becomes to optimize the expected reward conditioned on the context, namely $\mu_{\mathbf{x}}(\mathbf{C} = \mathbf{c}) = \mathbb{E}[Y \mid \text{do}(\mathbf{x}), \mathbf{C} = \mathbf{c}]$. Moreover, we assume that $\mathbf{C}$ is closed under ancestry, i.e., $An(\mathbf{C}) = \mathbf{C}$, so that context variables have only other context variables as ancestors and are therefore unaffected by interventions, thereby preserving their interpretation as pre-intervention information and avoiding time-ordering issues.

## 3 Possibly Optimal Arms for Causal Bandits

In this section, we revisit the results of Lee & Bareinboim (2019), which characterize possibly optimal arms in causal bandits when certain nodes in the causal graph are non-manipulable. Let $\mathbf{N} \subseteq \mathbf{V} \setminus \{Y\}$ denote the set of non-manipulable variables, noting that the reward variable $Y$ is also inherently non-manipulable.

**Definition 1.** *(**Unobserved Confounder (UC)-Territory** Lee & Bareinboim (2018)) Consider a causal graph $\mathcal{G}(\mathbf{V}, \mathbf{E})$ with reward node $Y$, and let $\mathcal{H}$ be the subgraph $\mathcal{G}[\mathsf{An}(Y)]$. A set of variables $\mathbf{T} \subseteq V(\mathcal{H})$ containing $Y$ is called a **UC-territory** on $\mathcal{G}$ with respect to $Y$ if $\mathsf{De}_{\mathcal{H}}(\mathbf{T}) = \mathbf{T}$ and $\mathsf{CC}_{\mathcal{H}}(\mathbf{T}) = \mathbf{T}$.*

A UC-territory is minimal if none of its proper subsets containing $Y$ is a UC-territory. A minimal UC-territory, denoted by $\mathsf{MUCT}(\mathcal{G}, Y)$, can be constructed by starting from the set $\{Y\}$ and alternately extending the current set by its c-component and its descendants until the set no longer changes.

**Definition 2.** ***(Interventional Border Lee & Bareinboim (2018))*** *Let $\mathbf{T}$ be a minimal UC-territory on $\mathcal{G}$ with respect to $Y$. Then, $\mathbf{X} = \mathsf{Pa}(\mathbf{T}) \setminus \mathbf{T}$ is called the interventional border of $\mathcal{G}$ with respect to $Y$, denoted by $\mathsf{IB}(\mathcal{G}, Y)$.*

**Lemma 1.** *Lee & Bareinboim (2018) For a graph $\mathcal{G}$ with reward node $Y$, $\mathsf{IB}(\mathcal{G}_{\overline{\mathbf{W}}}, Y)$ is a possibly optimal minimal intervention set (POMIS) for any $\mathbf{W} \subseteq \mathbf{V} \setminus \{Y\}$.*

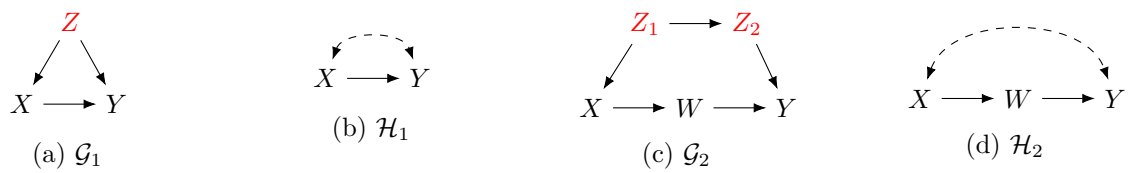



Figure 1: Original causal graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, along with their projections $\mathcal{H}_1$ and $\mathcal{H}_2$.

For a causal graph $\mathcal{G}$ in which all nodes except the reward node $Y$ are manipulable, we use the notation $\mathbb{P}_{\mathcal{G},Y}$ to denote the collection of all POMISs. In the more general setting where the variables in $\mathbf{N}$ cannot be intervened upon, we use $\mathbb{P}^{\mathbf{N}}_{\mathcal{G},Y}$ to denote the corresponding collection of possibly optimal minimal intervention sets. Following Lee & Bareinboim (2019), these intervention sets can be obtained through a projection step. We initialize a graph $\mathcal{H} = \langle \mathbf{V} \setminus \mathbf{N}, \emptyset \rangle$ and add a directed edge $V_i \to V_j$ if $V_i \to V_j$ is present in $\mathcal{G}$, or if there exists a directed path from $V_i$ to $V_j$ whose non-endpoint vertices all belong to $\mathbf{N}$. We add a bidirected edge $V_i \leftrightarrow V_j$ if this edge is present in $\mathcal{G}$, or if the projection of paths passing through variables in $\mathbf{N}$ induces latent confounding between $V_i$ and $V_j$. The results of Lee & Bareinboim (2019) show that the POMISs of the original causal bandit problem under the manipulability constraints can be obtained by enumerating the POMISs of the projected graph $\mathcal{H}$ using Lemma 1. Thus, $\mathbb{P}^{\mathbf{N}}_{\mathcal{G},Y} = \mathbb{P}_{\mathcal{H},Y}$.

Consider the causal graph $\mathcal{G}_1$ with reward node $Y$ and non-manipulable set $\mathbf{N} = \{Z\}$, and let $\mathcal{H}_1$ be the graph resulting from the projection step, as shown in Figure 1. The possibly optimal intervention sets for the original causal graph can be obtained from $\mathcal{H}_1$ using Lemma 1. In particular, $\mathbb{P}_{\mathcal{H}_1,Y} = \{\emptyset, \{X\}\} = \mathbb{P}^{\mathbf{N}}_{\mathcal{G}_1,Y}$. Therefore, the candidate optimal arms are $do()$, $do(X = 0)$, and $do(X = 1)$. Similarly, for the causal graph $\mathcal{G}_2$ with reward node $Y$ and non-manipulable set $\mathbf{N} = \{Z_1, Z_2\}$, the projected graph $\mathcal{H}_2$ is shown in Figure 1. In this case, $\mathbb{P}_{\mathcal{H}_2,Y} = \{\emptyset, \{W\}\} = \mathbb{P}^{\mathbf{N}}_{\mathcal{G}_2,Y}$, and the candidate optimal arms are $do()$, $do(W = 0)$, and $do(W = 1)$. More generally, we use $\mathcal{A}$ to denote the set of all interventions generated by all POMISs, that is, $\mathcal{A} := \bigcup_{\mathbf{S} \in \mathbb{P}^{\mathbf{N}}_{\mathcal{G},Y}} \{do(\mathbf{S} = \mathbf{s}) : \mathbf{s} \in \Omega(\mathbf{S})\}$, where $do(\emptyset)$ denotes the observational action.

We use the causal graph $\mathcal{G}_1$ to illustrate how samples obtained from different arms can be used to improve reward estimates across interventions. As an illustrative example, let $Z$ represent the age of a patient, let $X$ denote a medication or treatment, and let $Y$ represent the health outcome. The patient's age $Z$ has a causal effect on the outcome $Y$ but cannot be directly manipulated. The candidate optimal arms are $do()$, $do(X = 0)$, and $do(X = 1)$. Their expected rewards are given by $P(Y = 1 \mid do()) = \sum_{x,z} P(Y = 1 \mid x, z)P(x \mid z)P(z)$, $P(Y = 1 \mid do(X = 0)) = \sum_z P(Y = 1 \mid X = 0, z)P(z)$, and $P(Y = 1 \mid do(X = 1)) = \sum_z P(Y = 1 \mid X = 1, z)P(z)$. We assume that the causal graph is known, but the observational distribution is unknown. Since all non-intervened variables in the causal graph are observed after each interaction, samples collected under the different interventions can be combined to estimate the conditional distributions of variables given their parents. Because these conditional distributions are shared across the reward expressions of different interventions, observations obtained under one arm can improve the estimated rewards of the other arms. In the next section, we formalize this information sharing through a Bayesian posterior over the parameters of the observational distribution.

## 4 Bayesian Formulation for Causal Bandits

We consider a general probabilistic (Bayesian) formulation in which uncertain quantities are modeled as random variables. At each time $t$, the agent selects an action $A_t \in \mathcal{A}$ and observes a reward. For each action $a \in \mathcal{A}$ and time $t$, let $Y_{a,t} \in \mathcal{Y}$ denote the potential reward that would be observed at time $t$ if action $a$ were selected. There exists an unknown real-valued parameter vector $\boldsymbol{\theta}$ such that, conditioned on $\boldsymbol{\theta}$, the rewards are independent across time and satisfy $\mathbb{E}[Y_{A_t,t} \mid \boldsymbol{\theta}] = \mu_{\boldsymbol{\theta}}(A_t)$. The parameter $\boldsymbol{\theta}$ is random under a prior distribution, capturing uncertainty about the underlying reward-generating mechanism. Uncertainty in $\boldsymbol{\theta}$ induces uncertainty about the optimal action $A^\star \in \arg\max_{a\in\mathcal{A}} \mu_{\boldsymbol{\theta}}(a)$, where ties are broken using a fixed deterministic rule. The mean cumulative regret of a policy is defined as $\mathbb{E}[\mathrm{Reg}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \left(Y_{A^\star,t} - Y_{A_t,t}\right)\right]$, where the expectation is taken over the randomness in the actions $A_t$, the rewards, and the prior distribution over $\boldsymbol{\theta}$. This performance metric is commonly referred to as Bayesian regret or Bayesian risk.

Action $A_t$ is chosen based on the history of observations $\mathcal{F}_t = \left(\mathbf{C}_1, A_1, \mathbf{V}_{A_1,1}, Y_{A_1,1}, \ldots, \mathbf{C}_{t-1}, A_{t-1}, \mathbf{V}_{A_{t-1},t-1}, Y_{A_{t-1},t-1}\right)$ available before observing the context at time $t$. Formally, a randomized policy $\pi = (\pi_t)_{t\in\mathbb{N}}$ is a sequence of deterministic functions, where $\pi_t\left(\mathcal{F}_t, \mathbf{C}_t\right)$ specifies a probability distribution over the action set $\mathcal{A}$. Let $\mathcal{D}(\mathcal{A})$ denote the set of probability distributions over $\mathcal{A}$. After observing $\mathbf{C}_t$, the action $A_t$ is selected by sampling from $\pi_t\left(\mathcal{F}_t, \mathbf{C}_t\right)$. With some abuse of notation, we typically denote this distribution by $\pi_t$, where $\pi_t(a \mid \mathbf{c}_t) = \mathbb{P}\left(A_t = a \mid \mathcal{F}_t, \mathbf{C}_t = \mathbf{c}_t\right)$ denotes the probability assigned to action $a$ under the realized context $\mathbf{c}_t$. As shorthand notation, we use $P_t(\cdot)$ for $P(\cdot \mid \mathcal{F}_t)$ and $\mathbb{E}_t(\cdot)$ for $\mathbb{E}(\cdot \mid \mathcal{F}_t)$, respectively. After observing $\mathbf{C}_t = \mathbf{c}_t$, we define the context-conditioned posterior by $P_t^{\mathbf{c}_t}(\cdot) := P(\cdot \mid \mathcal{F}_t, \mathbf{C}_t = \mathbf{c}_t)$ and the corresponding conditional expectation by $\mathbb{E}_t^{\mathbf{c}_t}(\cdot)$.

In the contextual setting, at each time $t$, a context $\mathbf{C}_t \in \mathcal{C}$ is observed prior to selecting an action. The expected reward depends on both the action and the context and satisfies $\mathbb{E}[Y_{A_t,t} \mid \mathbf{C}_t = \mathbf{c}, \boldsymbol{\theta}] = \mu_{\boldsymbol{\theta}}(A_t, \mathbf{c})$. Accordingly, we define the oracle mapping $\Pi^\star : \mathcal{C} \to \mathcal{A}$ by $\Pi^\star(\mathbf{c}) := \arg\max_{a\in\mathcal{A}} \mu_{\boldsymbol{\theta}}(a, \mathbf{c})$, where ties are broken using a fixed deterministic rule. The mean cumulative Bayesian regret relative to this oracle mapping, conditional on the realized contexts $\{\mathbf{C}_t\}_{t=1}^T$, is given by

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(Y_{\Pi^\star(\mathbf{C}_t),t} - Y_{A_t,t}\right) \,\middle|\, \{\mathbf{C}_t\}_{t=1}^T\right]. \tag{1}$$

One remaining task is to specify the parameter $\boldsymbol{\theta}$ in the causal bandit setting. We assume that the causal graph contains no unobserved confounders. Under this assumption, any interventional distribution, including the conditional interventional distributions arising in the contextual setting, is identifiable from the observational distribution and the graph structure through the truncated factorization formula Pearl (2009). For instance, $P(\mathbf{v}_{\mathbf{V}\setminus\mathbf{X}} \mid do(\mathbf{x})) = \prod_{V_i\in\mathbf{V}\setminus\mathbf{X}} P(v_i \mid \mathbf{pa}_i)$, where the parent values are evaluated consistently with the intervention $\mathbf{X} = \mathbf{x}$. Thus, the observational distribution naturally determines the parameter $\boldsymbol{\theta}$.

Under the Markov assumption, the joint distribution factorizes as $P(\mathbf{v}) = \prod_{V_i\in\mathbf{V}} P(v_i \mid \mathbf{pa}_i)$. Rather than computing the posterior over the entire joint observational distribution, we compute the posterior over its local conditional probability distributions using the history of observations $\mathcal{F}_t$. For each variable $V_i \in \mathbf{V}$ with $r_i$ possible values $x_i^1, \ldots, x_i^{r_i}$, the local likelihood consists of a collection of multinomial distributions, one for each configuration of $\mathbf{Pa}_i$. Specifically, $P\left(V_i = x_i^k \mid \mathbf{Pa}_i = \mathbf{pa}_i^j, \boldsymbol{\theta}_{ij}\right) = \theta_{ijk}$, where $\theta_{ijk} > 0$ and $\sum_{k=1}^{r_i} \theta_{ijk} = 1$. Here, $\mathbf{pa}_i^1, \ldots, \mathbf{pa}_i^{q_i}$ denote the $q_i$ possible configurations of the parent set $\mathbf{Pa}_i$, where $q_i = \prod_{X_\ell\in\mathbf{Pa}_i} r_\ell$, and $\boldsymbol{\theta}_{ij} = (\theta_{ij1}, \ldots, \theta_{ijr_i})$ denotes the parameter vector associated with the $j$-th parent configuration. We define $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}, \ldots, \boldsymbol{\theta}_{iq_i})$, and the overall parameter vector is therefore $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$.

Similar to Heckerman et al. (2006), to ensure efficient computation, we assume parameter independence across the vectors $\boldsymbol{\theta}_{ij}$. Consequently, conditioned on the history of observations $\mathcal{F}_t$, each vector $\boldsymbol{\theta}_{ij}$ can be updated independently. Assuming that each vector $\boldsymbol{\theta}_{ij}$ follows a conjugate Dirichlet prior, namely $\mathrm{Dir}(\boldsymbol{\theta}_{ij} \mid \alpha_{ij1}, \ldots, \alpha_{ijr_i})$, the posterior distribution over the parameter space is given by

$$P_t(\boldsymbol{\theta}) = P\left(\boldsymbol{\theta} \mid \mathcal{F}_t\right) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} P_t(\boldsymbol{\theta}_{ij}). \tag{2}$$

**Algorithm 1:** Thompson Sampling for causal bandits with non-manipulable variables

**Input :** Causal graph $\mathcal{G}$ and the action set $\mathcal{A}$ induced by all POMISs of the projected graph.
**for** $t = 1$ **to** $T$ **do**
  Compute the posterior $P_t(\boldsymbol{\theta})$ using (2);
  Observe the realized context $\mathbf{C}_t = \mathbf{c}_t$;
  Compute the context-conditioned posterior $P_t^{\mathbf{c}_t}(\boldsymbol{\theta})$;
  Sample $\boldsymbol{\theta}_t \sim P_t^{\mathbf{c}_t}(\boldsymbol{\theta})$;
  Select action $a_t \leftarrow \arg\max_{a \in \mathcal{A}} \mathbb{E}[Y_{a,t} \mid \boldsymbol{\theta}_t, \mathbf{c}_t]$;
  Play action $a_t$, observe the reward and all non-intervened variables, and update the history $\mathcal{F}_{t+1}$;

$$P_t(\boldsymbol{\theta}_{ij}) = \mathrm{Dir}\left(\boldsymbol{\theta}_{ij} \mid \alpha_{ij1} + N_{ij1}(\mathcal{F}_t), \ldots, \alpha_{ijr_i} + N_{ijr_i}(\mathcal{F}_t)\right). \tag{3}$$

Here, $N_{ijk}(\mathcal{F}_t)$ denotes the number of times the realization $V_i = x_i^k$ has been observed in the history $\mathcal{F}_t$ when $\mathbf{Pa}_i = \mathbf{pa}_i^j$ and the corresponding action does not intervene on $V_i$. After observing the current context $\mathbf{C}_t = \mathbf{c}_t$, the posterior used for action selection is $P_t^{\mathbf{c}_t}(\boldsymbol{\theta}) = P(\boldsymbol{\theta} \mid \mathcal{F}_t, \mathbf{C}_t = \mathbf{c}_t)$, which is obtained by conditioning $P_t(\boldsymbol{\theta})$ on the observed context. Equation (2) can be used to compute the posterior over the parameter space and to draw posterior samples. Maintaining an accurate posterior is crucial for effective exploration and exploitation in the bandit problem.

## 5 Thompson Sampling Algorithm for Causal bandits

A bandit algorithm can select actions according to the posterior probability that an intervention is optimal, a strategy known as posterior sampling or Thompson sampling Thompson (1933). We propose a Thompson sampling algorithm for causal bandits with non-manipulable variables. Algorithm 1 maintains, at each round $t$, a posterior distribution over the parameter space, denoted by $P_t(\boldsymbol{\theta})$, capturing uncertainty given the history of observations up to round $t$. After observing the realized context $\mathbf{C}_t = \mathbf{c}_t$, the action at round $t$ is selected as $a_t \in \arg\max_{a \in \mathcal{A}} \mathbb{E}[Y_{a,t} \mid \boldsymbol{\theta}_t, \mathbf{c}_t]$, where $\boldsymbol{\theta}_t$ is an independent sample drawn from the context-conditioned posterior distribution $P_t(\boldsymbol{\theta} \mid \mathbf{C}_t = \mathbf{c}_t)$. The posterior is updated in a Bayesian manner after each interaction, consistent with the classical Thompson sampling framework. To analyze regret in the contextual setting, we introduce a nonnegative coefficient $\Gamma_t$ such that the expected instantaneous regret satisfies

$$\mathbb{E}_t\big[Y_{\Pi^\star(\mathbf{C}_t),t} - Y_{A_t,t} \mid \mathbf{C}_t\big] \leq \Gamma_t \sqrt{I_t(\Pi^\star(\mathbf{C}_t); (A_t, Y_{A_t,t}) \mid \mathbf{C}_t)}.$$

The quantity $\Gamma_t^2$ is referred to as the information ratio and characterizes the trade-off between incurring low regret and acquiring information about the optimal action under the realized context. The term $I_t(\Pi^\star(\mathbf{C}_t); (A_t, Y_{A_t,t}) \mid \mathbf{C}_t) := D_{\mathrm{KL}}(P_t(\Pi^\star(\mathbf{C}_t), A_t, Y_{A_t,t} \mid \mathbf{C}_t) \parallel P_t(\Pi^\star(\mathbf{C}_t) \mid \mathbf{C}_t)\, P_t(A_t, Y_{A_t,t} \mid \mathbf{C}_t))$ denotes the filtered mutual information between the optimal action under the realized context and the action-reward observation at round $t$. This quantity measures how much the observation collected at round $t$ reduces uncertainty about the optimal action under the realized context. Since $\Pi^\star(\mathbf{C}_t)$ is a deterministic function of the full oracle mapping $\Pi^\star$, the cumulative information gain is bounded by the entropy $H(\Pi^\star)$. Theorem 2 provides a bound on the mean cumulative regret.

**Theorem 2.** *The mean cumulative Bayesian regret of Algorithm 1 is bounded for any $T \in \mathbb{N}$ as follows, where $\Gamma \geq \Gamma_t$ for all $t \in \{1, \ldots, T\}$:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(Y_{\Pi^\star(\mathbf{C}_t),t} - Y_{A_t,t}\right)\right] \leq \Gamma \sqrt{H(\Pi^\star)\, T}.$$

The entropy $H(\Pi^\star)$ quantifies the decision-maker's initial uncertainty regarding the optimal-action mapping. The mean cumulative regret incurred by Algorithm 1 depends on the time horizon $T$, the entropy $H(\Pi^\star)$, and the worst-case upper bound on the information-ratio coefficient $\Gamma_t$. Following Russo & Van Roy (2016), we establish a simple worst-case upper bound given by $\Gamma_t^2 \leq \frac{|\mathcal{A}|}{2}$, or equivalently $\Gamma_t \leq \sqrt{\frac{|\mathcal{A}|}{2}}$. We adopt this

worst-case bound because the causal structure may provide no useful information sharing across interventions in the worst case. Therefore, the constant $\Gamma$ in Theorem 2 can be replaced with $\sqrt{\frac{|\mathcal{A}|}{2}}$, leading to the regret bound stated in Corollary 3. Moreover, since $\Pi^\star$ maps the context space $\mathcal{C}$ to the action space $\mathcal{A}$, the number of possible mappings is $|\mathcal{A}|^{|\mathcal{C}|}$ when $\mathcal{C}$ is finite, which implies that the entropy satisfies $H(\Pi^\star) \leq \log\left(|\mathcal{A}|^{|\mathcal{C}|}\right) = |\mathcal{C}| \log |\mathcal{A}|$.

**Corollary 3.** *Suppose that the context space $\mathcal{C}$ is finite. The mean cumulative Bayesian regret for Algorithm 1 is bounded for any $T \in \mathbb{N}$ as follows:*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(Y_{\Pi^\star(\mathbf{C}_t),t} - Y_{A_t,t}\right)\right] \leq \sqrt{\frac{T|\mathcal{A}||\mathcal{C}|\log|\mathcal{A}|}{2}}.$$

## 6 Information-Directed Sampling Algorithm for Causal bandits

Information-Directed Sampling (IDS) is a general framework for constructing decision-making algorithms. Rather than specifying a concrete sequence of computational steps, IDS defines an optimization criterion that guides the selection of actions. The framework balances two competing objectives: minimizing the expected instantaneous regret and acquiring informative observations about the identity of the optimal action. In the contextual setting, after observing the realized context $\mathbf{C}_t = \mathbf{c}_t$ at time $t$, IDS selects an action distribution that minimizes the ratio of the squared expected regret $\Delta_t(\pi \mid \mathbf{c}_t)^2$ to the information gain $g_t(\pi \mid \mathbf{c}_t)$ about the optimal action under that context, namely $\Pi^\star(\mathbf{c}_t)$, over all sampling distributions $\pi \in \mathcal{D}(\mathcal{A})$. The resulting IDS policy $\pi_t^{\mathrm{IDS}}(\mathbf{c}_t)$ is formally defined as:

$$\pi_t^{\mathrm{IDS}}(\mathbf{c}_t) \in \arg\min_{\pi \in \mathcal{D}(\mathcal{A})}\left\{\Psi_t(\pi \mid \mathbf{c}_t) := \frac{\Delta_t(\pi \mid \mathbf{c}_t)^2}{g_t(\pi \mid \mathbf{c}_t)}\right\}.$$

The expected instantaneous regret of a policy $\pi$ is defined as $\Delta_t(\pi \mid \mathbf{c}_t) = \sum_{a\in\mathcal{A}} \pi(a)\Delta_t(a \mid \mathbf{c}_t)$, where $\Delta_t(a \mid \mathbf{c}_t) := \mathbb{E}_t\left[Y_{\Pi^\star(\mathbf{c}_t),t} - Y_{a,t} \mid \mathbf{C}_t = \mathbf{c}_t\right]$ denotes the expected instantaneous regret incurred by selecting action $a$ at time step $t$ under context $\mathbf{c}_t$. Thus, the regret of a randomized policy is the probability-weighted average of the regrets of its constituent actions. Similarly, the information gain of a policy is defined as $g_t(\pi \mid \mathbf{c}_t) = \sum_{a\in\mathcal{A}} \pi(a) g_t(a \mid \mathbf{c}_t)$, where $g_t(a \mid \mathbf{c}_t) := I_t(\Pi^\star(\mathbf{c}_t); Y_{a,t} \mid \mathbf{C}_t = \mathbf{c}_t) = H_t(\Pi^\star(\mathbf{c}_t) \mid \mathbf{C}_t = \mathbf{c}_t) - H_t(\Pi^\star(\mathbf{c}_t) \mid \mathbf{C}_t = \mathbf{c}_t, Y_{a,t})$ denotes the information gained by selecting action $a$. In other words, $g_t(a \mid \mathbf{c}_t)$ measures the expected reduction in posterior uncertainty about the optimal action under context $\mathbf{c}_t$ after observing the corresponding reward $Y_{a,t}$. The quantity $\Psi_t(\pi \mid \mathbf{c}_t)$, referred to as the *information ratio*, quantifies the expected regret incurred per unit of information acquired and therefore captures the trade-off between exploiting actions with low immediate regret and exploring actions that provide information about the optimal action. At each time step, IDS greedily selects a policy that minimizes this criterion. We adopt the conventions $0/0 := 0$ and $x/0 := +\infty$ for every $x > 0$.

We use an example to illustrate why the quantity $g_t(a \mid \mathbf{c}_t)$ is useful for characterizing how informative a particular action is. Since we treat the observational distribution as defining the parameter space, one might argue that the empty intervention $do()$ is the most informative about the parameter $\boldsymbol{\theta}$, as it provides direct samples from the natural data-generating process. However, our objective is not merely to learn the full observational distribution, but rather to reduce uncertainty about the optimal action under the realized context $\mathbf{C}_t = \mathbf{c}_t$. Consider a simple causal graph in which $X \to Y$. The quantities relevant for identifying the optimal action are the context-dependent conditional probabilities $P(Y = 1 \mid X = 0, \mathbf{c}_t)$ and $P(Y = 1 \mid X = 1, \mathbf{c}_t)$. If $P(X = 0 \mid \mathbf{c}_t)$ is small relative to $P(X = 1 \mid \mathbf{c}_t)$, then samples drawn from the observational distribution will rarely provide information about the behavior of $Y$ when $X = 0$ under that context. In such a case, performing the intervention $do(X = 0)$ may be substantially more informative than $do()$ for determining the optimal action $\Pi^\star(\mathbf{c}_t)$. This distinction is precisely captured by $g_t(a \mid \mathbf{c}_t)$, which measures the expected reduction in posterior uncertainty about $\Pi^\star(\mathbf{c}_t)$ resulting from selecting action $a$. We now proceed to formally derive the expressions needed to compute $\Delta_t(a \mid \mathbf{c}_t)$ and $g_t(a \mid \mathbf{c}_t)$.

**Algorithm 2:** Information-Directed Sampling for Causal Bandits with Non-Manipulable Variables

**Input:** Causal graph $\mathcal{G}$, confidence level $\delta'$, posterior sample size $N$, and action set $\mathcal{A}$ induced by all POMISs of the projected graph

**for** $t = 1$ **to** $T$ **do**

- Compute the posterior $P_t(\boldsymbol{\theta})$ using (2);
- Observe the realized context $\mathbf{C}_t = \mathbf{c}_t$;
- Compute the context-conditioned posterior $P_t^{\mathbf{c}_t}(\boldsymbol{\theta}) = P(\boldsymbol{\theta} \mid \mathcal{F}_t, \mathbf{C}_t = \mathbf{c}_t)$;
- Draw $N$ independent samples from $P_t^{\mathbf{c}_t}(\boldsymbol{\theta})$;
- For every $a \in \mathcal{A}$, select any value $\widetilde{\Delta}_t(a \mid \mathbf{c}_t)$ from the confidence interval in Lemma 5, using $\delta = \frac{\delta'}{2T|\mathcal{A}|}$;
- For every $a \in \mathcal{A}$, select any value $\widetilde{g}_t(a \mid \mathbf{c}_t)$ from the confidence interval $\left[\underline{g}_t(a \mid \mathbf{c}_t), \overline{g}_t(a \mid \mathbf{c}_t)\right]$ in Lemma 6, using $\delta = \frac{\delta'}{2T|\mathcal{A}|}$;
- Set $\vec{\Delta}_t(\mathbf{c}_t) := [\widetilde{\Delta}_t(a \mid \mathbf{c}_t)]_{a \in \mathcal{A}}$;
- Set $\vec{g}_t(\mathbf{c}_t) := [\widetilde{g}_t(a \mid \mathbf{c}_t)]_{a \in \mathcal{A}}$;
- $a_t \leftarrow \text{IDSAction}\Big(\mathcal{A}, \vec{\Delta}_t(\mathbf{c}_t), \vec{g}_t(\mathbf{c}_t)\Big)$;
- Play action $a_t$, observe the reward and all non-intervened variables, and update the history $\mathcal{F}_{t+1}$;

**Lemma 4.** *The information gain associated with an action $a \in \mathcal{A}$ at time step $t$, conditional on the realized context $\mathbf{C}_t = \mathbf{c}_t$, is given by*

$$g_t(a \mid \mathbf{c}_t) = D_{\mathrm{KL}}(P_t^{\mathbf{c}_t}(\Pi^\star(\mathbf{c}_t), Y_a) \,\|\, P_t^{\mathbf{c}_t}(\Pi^\star(\mathbf{c}_t))\, P_t^{\mathbf{c}_t}(Y_a))\,. \tag{4}$$

*The corresponding posterior distributions satisfy*

$$P_t^{\mathbf{c}_t}(\Pi^\star(\mathbf{c}_t) = a^\star) = \mathbb{E}_{\boldsymbol{\theta} \sim P_t^{\mathbf{c}_t}(\boldsymbol{\theta})}\left[\mathbb{1}\left\{a^\star = \arg\max_{a' \in \mathcal{A}} \mu_{\boldsymbol{\theta}}(a', \mathbf{c}_t)\right\}\right], \tag{5}$$

$$P_t^{\mathbf{c}_t}(Y_a = y) = \mathbb{E}_{\boldsymbol{\theta} \sim P_t^{\mathbf{c}_t}(\boldsymbol{\theta})}\left[P(Y_a = y \mid \boldsymbol{\theta}, \mathbf{c}_t)\right], \tag{6}$$

*and*

$$\begin{aligned} &P_t^{\mathbf{c}_t}(\Pi^\star(\mathbf{c}_t) = a^\star, Y_a = y) \\ &\quad = \mathbb{E}_{\boldsymbol{\theta} \sim P_t^{\mathbf{c}_t}(\boldsymbol{\theta})}\left[\mathbb{1}\left\{a^\star = \arg\max_{a' \in \mathcal{A}} \mu_{\boldsymbol{\theta}}(a', \mathbf{c}_t)\right\} P(Y_a = y \mid \boldsymbol{\theta}, \mathbf{c}_t)\right], \end{aligned} \tag{7}$$

*for every $a^\star \in \mathcal{A}$ and $y \in \{0, 1\}$.*

*Moreover, the expected instantaneous regret associated with action $a$ at time step $t$ under context $\mathbf{c}_t$ is*

$$\begin{aligned} \Delta_t(a \mid \mathbf{c}_t) = \mathbb{E}_{\boldsymbol{\theta} \sim P_t^{\mathbf{c}_t}(\boldsymbol{\theta})}\Bigg[&\sum_{a^\star \in \mathcal{A}} \mathbb{1}\left\{a^\star = \arg\max_{a' \in \mathcal{A}} \mu_{\boldsymbol{\theta}}(a', \mathbf{c}_t)\right\} \\ &\times P(Y_{a^\star} = 1 \mid \boldsymbol{\theta}, \mathbf{c}_t) - P(Y_a = 1 \mid \boldsymbol{\theta}, \mathbf{c}_t)\Bigg]. \end{aligned} \tag{8}$$

When the parameter space is finite and discrete, the quantities $g_t(a \mid \mathbf{c}_t)$ and $\Delta_t(a \mid \mathbf{c}_t)$ can be evaluated exactly by summing over all possible values of $\boldsymbol{\theta}$. In our contextual causal bandit setting, however, $\boldsymbol{\theta}$ parameterizes the observational distribution and belongs to a continuous probability simplex. Consequently, the posterior expectations defining $g_t(a \mid \mathbf{c}_t)$ and $\Delta_t(a \mid \mathbf{c}_t)$ are generally analytically intractable. We therefore approximate these quantities using Monte Carlo samples drawn from the context-conditioned posterior distribution $P_t^{\mathbf{c}_t}(\boldsymbol{\theta})$. The following results control the approximation error introduced by this sampling procedure.

**Lemma 5** (Concentration of the Expected Regret)**.** *Fix a time step $t$, a realized context $\mathbf{C}_t = \mathbf{c}_t$, an action $a \in \mathcal{A}$, and $\delta \in (0,1)$. Let $\widehat{\Delta}_t(a \mid \mathbf{c}_t)$ be the Monte Carlo estimator of $\Delta_t(a \mid \mathbf{c}_t)$ computed using $N$ independent samples from $P_t^{\mathbf{c}_t}(\boldsymbol{\theta})$. Then, with probability at least $1-\delta$,*

$$\left|\widehat{\Delta}_t(a \mid \mathbf{c}_t) - \Delta_t(a \mid \mathbf{c}_t)\right| \leq \sqrt{\frac{2}{N}\log\left(\frac{2}{\delta}\right)}.$$

To control the Monte Carlo approximation error in the information gain, we derive simultaneous concentration bounds for the posterior probabilities $P_t^{\mathbf{c}_t}(\Pi^\star(\mathbf{c}_t) = a^\star)$, $P_t^{\mathbf{c}_t}(Y_a = y)$, and $P_t^{\mathbf{c}_t}(\Pi^\star(\mathbf{c}_t) = a^\star, Y_a = y)$. These bounds yield computable lower and upper confidence bounds for $g_t(a \mid \mathbf{c}_t)$ that hold uniformly over all actions.

**Lemma 6** (Concentration of the Information Gain)**.** *Fix a time step $t$, a realized context $\mathbf{C}_t = \mathbf{c}_t$, and $\delta \in (0,1)$. For each $a \in \mathcal{A}$, let $\underline{g}_t(a \mid \mathbf{c}_t)$ and $\overline{g}_t(a \mid \mathbf{c}_t)$ denote the lower and upper confidence bounds defined in* (35) *and* (33)*, respectively, and constructed using $N$ independent samples from the context-conditioned posterior distribution $P_t^{\mathbf{c}_t}(\boldsymbol{\theta})$. Then, with probability at least $1-\delta$ over the Monte Carlo samples,*

$$\underline{g}_t(a \mid \mathbf{c}_t) \leq g_t(a \mid \mathbf{c}_t) \leq \overline{g}_t(a \mid \mathbf{c}_t), \qquad \forall a \in \mathcal{A}.$$

The proof of Lemma 6, together with the explicit construction of the confidence bounds, is provided in Appendix A.6.

**Theorem 7.** *Fix $\delta' \in (0,1)$ and a horizon $T \in \mathbb{N}$. With probability at least $1-\delta'$ over the Monte Carlo samples used by Algorithm 2, the Bayesian regret satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(Y_{\Pi^\star(\mathbf{C}_t),t} - Y_{A_t,t}\right)\right] \leq \sqrt{H(\Pi^\star)\left(\frac{T|\mathcal{A}|}{2} + 2\sum_{t=1}^{T}\gamma_t\right)}. \tag{9}$$

*Here, for each time step $t$,*

$$\gamma_t := \sup_{\substack{\pi \in S_{|\mathcal{A}|},\\ \boldsymbol{\Delta}_t^1(\mathbf{c}_t), \boldsymbol{\Delta}_t^2(\mathbf{c}_t) \in \mathcal{C}_t^{\Delta}(\mathbf{c}_t),\\ \boldsymbol{g}_t^1(\mathbf{c}_t), \boldsymbol{g}_t^2(\mathbf{c}_t) \in \mathcal{C}_t^{g}(\mathbf{c}_t)}} \left|\frac{\left(\pi^\top \boldsymbol{\Delta}_t^1(\mathbf{c}_t)\right)^2}{\pi^\top \boldsymbol{g}_t^1(\mathbf{c}_t)} - \frac{\left(\pi^\top \boldsymbol{\Delta}_t^2(\mathbf{c}_t)\right)^2}{\pi^\top \boldsymbol{g}_t^2(\mathbf{c}_t)}\right|, \tag{10}$$

*where*

$$S_{|\mathcal{A}|} := \left\{\pi \in \mathbb{R}_+^{|\mathcal{A}|} : \sum_{a \in \mathcal{A}} \pi(a) = 1\right\}.$$

*The vectors $\boldsymbol{\Delta}_t^1(\mathbf{c}_t)$ and $\boldsymbol{\Delta}_t^2(\mathbf{c}_t)$ range over the componentwise confidence set $(\mathcal{C}_t^{\Delta}(\mathbf{c}_t))$ obtained from Lemma 5, while $\boldsymbol{g}_t^1(\mathbf{c}_t)$ and $\boldsymbol{g}_t^2(\mathbf{c}_t)$ range over the componentwise confidence set $(\mathcal{C}_t^{g}(\mathbf{c}_t))$ obtained from Lemma 6, all evaluated at the realized context $\mathbf{C}_t = \mathbf{c}_t$. Both confidence sets are constructed using $\delta = \frac{\delta'}{2T|\mathcal{A}|}$. The expectation in* (9) *is taken over the prior, contexts, rewards, and action-selection randomness, conditional on the Monte Carlo confidence event.*

We analyze the regret of the proposed IDS algorithm for causal bandits with non-manipulable variables (Algorithm 2). Theorem 7 provides a Bayesian regret bound in which the additional term involving $\gamma_t$ quantifies the effect of computing the information ratio using Monte Carlo estimates. The widths of the confidence intervals in Lemma 5 and Lemma 6 scale as $O\left(\sqrt{\log(c/\delta)/N}\right)$ for an appropriate constant $c$. Therefore, increasing the posterior sample size $N$ shrinks the confidence sets and can only decrease, or leave unchanged, $\gamma_t$. Similarly, decreasing $\delta$ produces a higher-confidence guarantee but wider confidence intervals, which can increase $\gamma_t$, whereas increasing $\delta$ produces narrower intervals and can decrease $\gamma_t$. In Theorem 7, the choice $\delta = \delta'/(2T|\mathcal{A}|)$ balances the simultaneous confidence requirement across all rounds and actions with the resulting estimation error. Provided that the relevant information-gain denominators remain bounded away from zero, the confidence sets contract as $N$ increases, and consequently $\gamma_t$ approaches zero. Thus, $\gamma_t$ measures the worst-case inflation of the information ratio caused by finite Monte Carlo estimation. In the idealized oracle setting, the information ratio is computed exactly, the confidence sets collapse to the true values, and $\gamma_t = 0$ for every $t$.

**Corollary 8** (IDS with Oracle Access)**.** *Suppose the information ratio can be computed exactly at every round, equivalently, the Monte Carlo estimation error vanishes so that $\gamma_t = 0$ for all $t$. Then, for any $T \in \mathbb{N}$, the mean cumulative regret of Algorithm 2 satisfies*

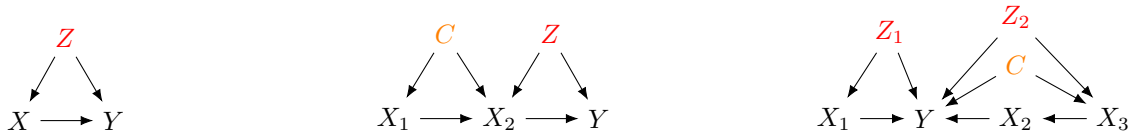
$$\mathbb{E}\left[\sum_{t=1}^{T}\left(Y_{\Pi^\star(\mathbf{C}_t),t} - Y_{A_t,t}\right)\right] \leq \sqrt{\frac{|\mathcal{A}|}{2}\, T\, H(\Pi^\star)}.$$

This result follows directly from the standard IDS analysis when the information ratio is computed exactly. To illustrate how causal structure can affect the prior distribution over optimal interventions and thereby tighten the regret bound, consider the causal graph in Figure 2(a), consisting of the binary nodes $X$, $Y$, and $Z$, with edges $X \to Y$, $Z \to X$, and $Z \to Y$, where $Y$ is the reward node and $Z$ is non-manipulable. There is no context variable in this example. Because $Z$ cannot be manipulated, the feasible intervention set is $\mathcal{A} = \{do(), do(X = 1), do(X = 0)\}$. Unlike an unstructured bandit model in which the arms are assigned unrelated reward parameters, the expected rewards of these interventions are jointly determined by the same causal mechanisms and shared conditional probability tables. Thus, the causal model induces a joint prior over the intervention rewards and, consequently, a prior distribution over the optimal intervention $\Pi^\star$. Assuming no prior observations, we place independent $\text{Beta}(1, 1)$ priors on the Bernoulli parameters in each row of the conditional probability tables and estimate $P(\Pi^\star = a)$ using $N = 50{,}000$ Monte Carlo samples. We obtain $P(\Pi^\star = do()) \approx 0.08$, $P(\Pi^\star = do(X = 1)) \approx 0.46$, and $P(\Pi^\star = do(X = 0)) \approx 0.46$. By Hoeffding's inequality and a union bound over the three interventions, these estimates have a simultaneous error bound of approximately $\pm 0.0093$ at overall confidence level 0.999. The resulting entropy is $H(\Pi^\star) = -\sum_{a\in\mathcal{A}} P(\Pi^\star = a) \log P(\Pi^\star = a) \approx 0.9165$, which is strictly smaller than the worst-case value $\log |\mathcal{A}| = \log 3 \approx 1.0986$. Consequently, the oracle IDS regret bound becomes $\mathbb{E}[\text{Reg}_T] \leq \sqrt{3H(\Pi^\star)T/2} \approx 1.1725\sqrt{T}$, whereas the worst-case uniform-prior benchmark is $\sqrt{3\log(3)T/2} \approx 1.2837\sqrt{T}$. Thus, in this example, the causal model reduces the leading regret coefficient by approximately 8.7% relative to the worst-case entropy bound. Moreover, because the interventions share the same causal mechanisms, observations collected under one intervention can update beliefs about the rewards and information gains of other interventions, although this additional cross-intervention information sharing is not explicitly captured by the worst-case factor $|\mathcal{A}|/2$.

## 7 Experiments

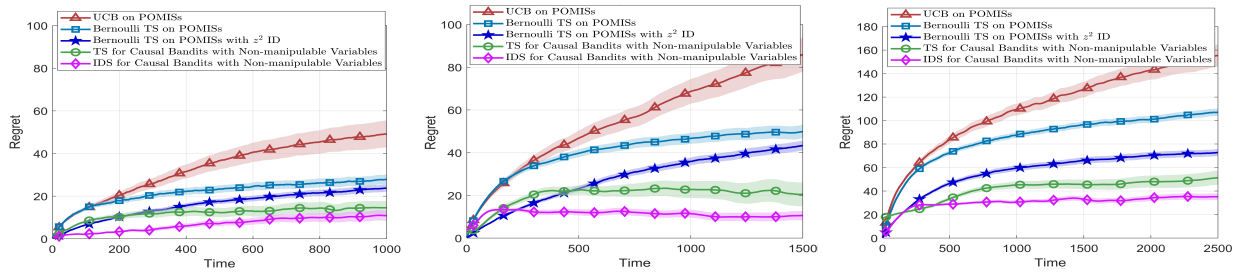
We empirically compare the performance of our proposed algorithms with existing baselines on three tasks of increasing complexity. The corresponding causal graphs are shown in Figure 2, where red nodes denote non-manipulable variables, $C$ denotes a context variable, and $Y$ denotes the reward variable. For **Task 1** (Figure 2(a)), the set of POMISs after the projection step is $\{\emptyset, \{X\}\}$, resulting in the possibly optimal arms $do()$, $do(X = 0)$, and $do(X = 1)$. For **Task 2** (Figure 2(b)), the set of POMISs is $\{\{X_1\}, \{X_2\}\}$, yielding the possibly optimal arms $do(X_1 = 0)$, $do(X_1 = 1)$, $do(X_2 = 0)$, and $do(X_2 = 1)$. For **Task 3** (Figure 2(c)), the set of POMISs is $\{\emptyset, \{X_1\}, \{X_2\}, \{X_1, X_2\}\}$, resulting in a total of nine possibly optimal arms.

We evaluate three baseline methods, including two non-causal algorithms: vanilla UCB and Bernoulli Thompson Sampling (TS), both applied to the set of possibly optimal interventions identified from the POMISs. The third baseline is the algorithm proposed in Lee & Bareinboim (2019), which uses $z^2$ID to construct multiple estimators of each arm's reward distribution and combines them using minimum-variance convex weights obtained through quadratic programming. For all three tasks, the CPTs are randomly sampled to construct Bayesian networks consistent with the corresponding causal graphs. For each baseline and each proposed algorithm, we run 500 independent trials and plot the average cumulative regret, together with bands corresponding to two standard deviations, in Figure 3. For tasks containing a context variable, a new context value is sampled independently at each round, and all algorithms select actions conditioned on the observed context. For IDS, the number of posterior samples used to estimate the expected regret and information gain is set to $N = 1000$. The results show that both proposed algorithms outperform the baselines by more effectively exploiting the causal structure and the information shared across interventions. In particular, IDS outperforms Thompson Sampling by explicitly balancing expected instantaneous regret against information gain, thereby favoring informative interventions that reduce posterior uncertainty about the context-dependent optimal action.



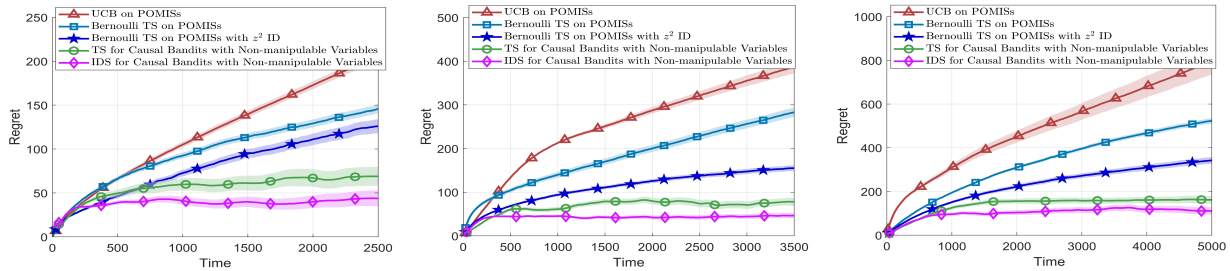

(a) Task 1 (b) Task 2 (c) Task 3

Figure 2: Causal graphs used in the experiments




(a) Task 1 (b) Task 2 (c) Task 3

Figure 3: Regret versus time for the three tasks with the corresponding causal graphs shown in Figure 2




(a) $N_{\mathcal{G}} = 10, \rho = 0.1$ (b) $N_{\mathcal{G}} = 15, \rho = 0.1$ (c) $N_{\mathcal{G}} = 15, \rho = 0.2$

Figure 4: Regret versus time for Erdős–Rényi random chordal graphs with varying graph density

We further compare our proposed algorithms with the baseline methods on Erdős–Rényi random chordal graphs with varying numbers of nodes $N_{\mathcal{G}}$ and graph densities $\rho$. After sampling each causal graph, 15% of its nodes are randomly designated as non-manipulable, and a subset of the remaining nodes is designated as context variables. The results are shown in Figure 4, and the graph-generation procedure is described in Section A.10 of the supplementary material. As in the three structured tasks, the proposed algorithms achieve lower cumulative regret than the baselines on the randomly generated graphs. Finally, we consider an additional synthetic causal graph motivated by a real-world healthcare scenario in Section A.9 of the supplementary material, further illustrating the applicability of the proposed framework.

## 8 Conclusion

We study contextual causal bandits with non-manipulable variables and introduce Bayesian Thompson Sampling (TS) and Information-Directed Sampling (IDS) algorithms for this setting. By maintaining a posterior distribution over the conditional probability tables of the causal model, the proposed methods exploit information shared across interventions to improve learning efficiency. We establish an entropy-dependent sublinear Bayesian regret bound for Thompson Sampling. For IDS, we derive a regret guarantee that explicitly captures the additional error introduced by Monte Carlo approximation, while the oracle version recovers the standard sublinear IDS bound. Empirical results on structured and randomly generated causal graphs show that the proposed algorithms outperform existing causal and non-causal baselines. Overall, our framework provides a principled approach for sequential decision-making in causal systems containing variables that influence the reward but cannot be directly manipulated.

## References


Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pp. 359–376. JMLR Workshop and Conference Proceedings, 2011.

David Heckerman, Christopher Meek, and Gregory Cooper. A bayesian approach to causal discovery. *Innovations in Machine Learning: Theory and Applications*, pp. 1–28, 2006.

Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th annual conference on information sciences and systems (CISS)*, pp. 1–6. IEEE, 2014.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pp. 423–439. PMLR, 2014.

Kwang-Sung Jun and Chicheng Zhang. Crush optimism with pessimism: Structured bandits beyond asymptotic optimality. *Advances in Neural Information Processing Systems*, 33:6366–6376, 2020.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

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

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 2573–2583. Curran Associates, Inc., 2018.

Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4164–4172, 2019.

Adam J Mersereau, Paat Rusmevichientong, and John N Tsitsiklis. A structured multiarmed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, 54(12):2787–2802, 2009.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Muhammad Qasim Elahi, Mahsa Ghasemi, and Murat Kocaoglu. Partial structure discovery is sufficient for no-regret learning in causal bandits. *arXiv e-prints*, pp. arXiv–2411, 2024.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in neural information processing systems*, 27, 2014.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

Eric Schulz, Nicholas T Franklin, and Samuel J Gershman. Finding structure in multi-armed bandits. *Cognitive psychology*, 119:101261, 2020.

Rajat Sen, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Identifying best interventions through online importance sampling. *arXiv preprint arXiv:1701.02789*, 2017.

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Chandler Squires, Sara Magliacane, Kristjan Greenewald, Dmitriy Katz, Murat Kocaoglu, and Karthikeyan Shanmugam. Active structure learning of causal dags via directed clique trees. *Advances in Neural Information Processing Systems*, 33:21500–21511, 2020.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pp. 567–573, 2002.

Andrea Tirinzoni, Alessandro Lazaric, and Marcello Restelli. A novel confidence-based algorithm for structured bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3175–3185. PMLR, 2020.

Bart Van Parys and Negin Golrezaei. Optimal learning for structured bandits. *Management Science*, 70(6): 3951–3998, 2024.

Joannes Vermorel and Mehryar Mohri. Multi-armed bandit algorithms and empirical evaluation. In *European conference on machine learning*, pp. 437–448. Springer, 2005.

Runzhe Wan, Lin Ge, and Rui Song. Towards scalable and robust structured bandits: A meta-learning framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 1144–1173. PMLR, 2023.

Lai Wei, Muhammad Qasim Elahi, Mahsa Ghasemi, and Murat Kocaoglu. Approximate allocation matching for structural causal bandits with unobserved confounders. *Advances in Neural Information Processing Systems*, 36, 2024.

Akihiro Yabe, Daisuke Hatano, Hanna Sumita, Shinji Ito, Naonori Kakimura, Takuro Fukunaga, and Ken-Ichi Kawarabayashi. Causal bandits with propagating inference. *arXiv preprint arXiv:1806.02252*, 2018.

# A Information-Directed Sampling for Causal Bandits (Supplementary Material)

## A.1 Optimizing the Contextual Information Ratio

If $\vec{g}_t(\mathbf{c}_t) = \vec{0}$, then none of the available actions provides information about the optimal action under context $\mathbf{c}_t$. In this case, we select an action that minimizes the expected instantaneous regret. Otherwise, at time $t$ we select an action by solving

$$\min_{\pi \in S_{|\mathcal{A}|}} \frac{\left(\pi^\top \vec{\Delta}_t(\mathbf{c}_t)\right)^2}{\pi^\top \vec{g}_t(\mathbf{c}_t)}, \tag{11}$$

where $S_{|\mathcal{A}|} = \{\pi \in \mathbb{R}_+^{|\mathcal{A}|} : \sum_k \pi_k = 1\}$ is the probability simplex. We use the conventions $0/0 := 0$ and $x/0 := +\infty$ for $x > 0$.

Russo and Van Roy Russo & Van Roy (2014) show that this optimization problem is convex and admits an optimal solution with at most two nonzero components. Therefore, although IDS is a randomized policy, it suffices to randomize over at most two actions at each round.

**Algorithm 3:** Function to optimize the contextual information ratio

**Function** `IDSAction`$(\mathcal{A}, \vec{\Delta}_t(\mathbf{c}_t), \vec{g}_t(\mathbf{c}_t))$**:**

  **if** $\vec{g}_t(\mathbf{c}_t) = \vec{0}$ **then**

    **return** an action $a^* \in \arg\min_{a \in \mathcal{A}} \Delta_t(a \mid \mathbf{c}_t)$.

  For each pair $a, a' \in \mathcal{A}$ such that $a \neq a'$ compute

$$q_{a,a'} \leftarrow \arg\min_{q \in [0,1]} \frac{\left(q\,\Delta_t(a \mid \mathbf{c}_t) + (1-q)\,\Delta_t(a' \mid \mathbf{c}_t)\right)^2}{q\,g_t(a \mid \mathbf{c}_t) + (1-q)\,g_t(a' \mid \mathbf{c}_t)}.$$

  Select

$$(a^*, a^{**}) \leftarrow \arg\min_{a \neq a'} \frac{\left(q_{a,a'}\,\Delta_t(a \mid \mathbf{c}_t) + (1-q_{a,a'})\,\Delta_t(a' \mid \mathbf{c}_t)\right)^2}{q_{a,a'}\,g_t(a \mid \mathbf{c}_t) + (1-q_{a,a'})\,g_t(a' \mid \mathbf{c}_t)}.$$

  **return** $a^*$ with probability $q_{a^*,a^{**}}$ and $a^{**}$ with probability $1 - q_{a^*,a^{**}}$.

**End Function**

## A.2 Proof of Theorem 2

$$\mathbb{E}[\mathrm{Reg}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \left(Y_{\Pi^\star(\mathbf{C}_t),t} - Y_{A_t,t}\right)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_t\left[Y_{\Pi^\star(\mathbf{C}_t),t} - Y_{A_t,t} \mid \mathbf{C}_t\right]\right]. \tag{12}$$

We introduce a nonnegative constant $\Gamma_t$ such that the per-round regret satisfies

$$\mathbb{E}_t\left[Y_{\Pi^\star(\mathbf{C}_t),t} - Y_{A_t,t} \mid \mathbf{C}_t\right] \leq \Gamma_t \sqrt{I_t(\Pi^\star(\mathbf{C}_t); (A_t, Y_{A_t,t}) \mid \mathbf{C}_t)}. \tag{13}$$

Then,

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_T] &\leq \mathbb{E}\left[\sum_{t=1}^{T} \Gamma_t \sqrt{I_t(\Pi^\star(\mathbf{C}_t); (A_t, Y_{A_t,t}) \mid \mathbf{C}_t)}\right] \\
&\leq \Gamma\, \mathbb{E}\left[\sum_{t=1}^{T} \sqrt{I_t(\Pi^\star(\mathbf{C}_t); (A_t, Y_{A_t,t}) \mid \mathbf{C}_t)}\right] \\
&\leq \Gamma \sqrt{T\, \mathbb{E}\left[\sum_{t=1}^{T} I_t(\Pi^\star(\mathbf{C}_t); (A_t, Y_{A_t,t}) \mid \mathbf{C}_t)\right]}.
\end{aligned}$$

The second inequality follows from the assumption that $\Gamma \geq \Gamma_t$ for all $t \in \mathbb{N}$, and the third inequality follows from the Cauchy–Schwarz and Jensen inequalities.

Since $\Pi^\star(\mathbf{C}_t)$ is a deterministic function of the full optimal-action mapping $\Pi^\star$, the data-processing inequality gives

$$I_t(\Pi^\star(\mathbf{C}_t); (A_t, Y_{A_t,t}) \mid \mathbf{C}_t) \leq I_t(\Pi^\star; (A_t, Y_{A_t,t}) \mid \mathbf{C}_t)\,. \tag{14}$$

Let $\mathcal{O}_t$ denote the complete observation collected at time $t$, including $\mathbf{C}_t$, $A_t$, $Y_{A_t,t}$, and all observed non-intervened variables, so that $\mathcal{F}_{t+1} = \sigma(\mathcal{F}_t, \mathcal{O}_t)$. By the chain rule for mutual information,

$$I_t(\Pi^\star; (A_t, Y_{A_t,t}) \mid \mathbf{C}_t) \leq I_t(\Pi^\star; \mathcal{O}_t)\,. \tag{15}$$

By the definition of conditional mutual information,

$$I_t(\Pi^\star; \mathcal{O}_t) = H(\Pi^\star \mid \mathcal{F}_t) - \mathbb{E}[H(\Pi^\star \mid \mathcal{F}_{t+1}) \mid \mathcal{F}_t]\,.$$

Summing over $t$ yields a telescoping series:

$$\mathbb{E}\left[\sum_{t=1}^{T} I_t(\Pi^\star(\mathbf{C}_t); (A_t, Y_{A_t,t}) \mid \mathbf{C}_t)\right] \leq H(\Pi^\star) - \mathbb{E}[H(\Pi^\star \mid \mathcal{F}_{T+1})] \leq H(\Pi^\star). \tag{16}$$

Substituting this bound gives

$$\mathbb{E}[\mathrm{Reg}_T] \leq \Gamma\sqrt{H(\Pi^\star)\,T}. \tag{17}$$

This completes the proof of Theorem 2.

### A.3 Proof of Corollary 3

We fix a realization of $\mathcal{F}_t$ and condition on the realized context $\mathbf{C}_t = \mathbf{c}$. Write $p_a := P_t(\Pi^\star(\mathbf{c}) = a \mid \mathbf{C}_t = \mathbf{c})$. By the TS posterior-matching property, $P_t(A_t = a \mid \mathbf{C}_t = \mathbf{c}) = p_a$. Moreover, conditional on $\mathcal{F}_t$ and $\mathbf{C}_t = \mathbf{c}$, the action $A_t$ and the optimal action $\Pi^\star(\mathbf{c})$ are independent because $A_t$ is generated using an independent posterior sample. Also let $K = |\mathcal{A}|$ be the total number of actions.

Define the posterior mean reward of action $a$ at context $\mathbf{c}$:

$$\bar{\mu}_a := E_t[Y_{a,t} \mid \mathbf{C}_t = \mathbf{c}].$$

Also define the posterior mean reward under the optimal action for that context:

$$\bar{\mu}^\star := E_t[Y_{\Pi^\star(\mathbf{c}),t} \mid \mathbf{C}_t = \mathbf{c}] = \sum_{a\in\mathcal{A}} p_a\, E_t[Y_{a,t} \mid \Pi^\star(\mathbf{c}) = a, \mathbf{C}_t = \mathbf{c}].$$

Then the conditional instantaneous regret is

$$\begin{aligned}\Delta_t(\mathbf{c}) &:= E_t\big[Y_{\Pi^\star(\mathbf{c}),t} - Y_{A_t,t} \mid \mathbf{C}_t = \mathbf{c}\big] \\ &= \bar{\mu}^\star - \sum_{a\in\mathcal{A}} P_t(A_t = a \mid \mathbf{C}_t = \mathbf{c})\,\bar{\mu}_a \\ &= \sum_{a\in\mathcal{A}} p_a\,(E_t[Y_{a,t} \mid \Pi^\star(\mathbf{c}) = a, \mathbf{C}_t = \mathbf{c}] - \bar{\mu}_a)\,.\end{aligned} \tag{18}$$

Applying the Cauchy–Schwarz inequality gives

$$\Delta_t(\mathbf{c})^2 \leq K \sum_{a\in\mathcal{A}} p_a^2\,(E_t[Y_{a,t} \mid \Pi^\star(\mathbf{c}) = a, \mathbf{C}_t = \mathbf{c}] - \bar{\mu}_a)^2\,. \tag{19}$$

Because rewards are Bernoulli, for each $a$ the two distributions

$$P_t(Y_{a,t} \in \cdot \mid \Pi^\star(\mathbf{c}) = a, \mathbf{C}_t = \mathbf{c}) \quad \text{and} \quad P_t(Y_{a,t} \in \cdot \mid \mathbf{C}_t = \mathbf{c})$$

are distributions on $\{0, 1\}$, and their total variation distance equals the absolute difference of their means:

$$TV(P_t(Y_{a,t} \mid \Pi^\star(\mathbf{c}) = a, \mathbf{C}_t = \mathbf{c}),\, P_t(Y_{a,t} \mid \mathbf{C}_t = \mathbf{c})) = |E_t[Y_{a,t} \mid \Pi^\star(\mathbf{c}) = a, \mathbf{C}_t = \mathbf{c}] - \bar{\mu}_a|\,.$$

Applying Pinsker's inequality for each $a$ gives

$$(E_t[Y_{a,t} \mid \Pi^\star(\mathbf{c}) = a, \mathbf{C}_t = \mathbf{c}] - \bar{\mu}_a)^2 \leq \frac{1}{2} KL(P_t(Y_{a,t} \mid \Pi^\star(\mathbf{c}) = a, \mathbf{C}_t = \mathbf{c}) \parallel P_t(Y_{a,t} \mid \mathbf{C}_t = \mathbf{c}))\,. \tag{20}$$

Plugging (20) into (19) yields

$$\Delta_t(\mathbf{c})^2 \leq \frac{K}{2} \sum_{a \in \mathcal{A}} p_a^2\, KL(P_t(Y_{a,t} \mid \Pi^\star(\mathbf{c}) = a, \mathbf{C}_t = \mathbf{c}) \parallel P_t(Y_{a,t} \mid \mathbf{C}_t = \mathbf{c}))\,. \tag{21}$$

We now relate the right-hand side to the information obtained under Thompson Sampling. Since $A_t$ and $\Pi^\star(\mathbf{c})$ are conditionally independent given $\mathcal{F}_t$ and $\mathbf{C}_t = \mathbf{c}$, we have

$$\begin{aligned}
&I_t(\Pi^\star(\mathbf{c}); (A_t, Y_{A_t,t}) \mid \mathbf{C}_t = \mathbf{c}) \\
&= \sum_{a \in \mathcal{A}} p_a\, I_t(\Pi^\star(\mathbf{c}); Y_{a,t} \mid \mathbf{C}_t = \mathbf{c}) \\
&= \sum_{a \in \mathcal{A}} \sum_{a' \in \mathcal{A}} p_a p_{a'}\, KL(P_t(Y_{a,t} \mid \Pi^\star(\mathbf{c}) = a', \mathbf{C}_t = \mathbf{c}) \parallel P_t(Y_{a,t} \mid \mathbf{C}_t = \mathbf{c})) \\
&\geq \sum_{a \in \mathcal{A}} p_a^2\, KL(P_t(Y_{a,t} \mid \Pi^\star(\mathbf{c}) = a, \mathbf{C}_t = \mathbf{c}) \parallel P_t(Y_{a,t} \mid \mathbf{C}_t = \mathbf{c}))\,.
\end{aligned} \tag{22}$$

Combining (21) and (22) gives

$$\Delta_t(\mathbf{c})^2 \leq \frac{K}{2} I_t(\Pi^\star(\mathbf{c}); (A_t, Y_{A_t,t}) \mid \mathbf{C}_t = \mathbf{c})\,.$$

Therefore, for every realized context $\mathbf{c}$,

$$\Gamma_t^2 \leq \frac{K}{2},$$

or equivalently,

$$\Gamma_t \leq \sqrt{\frac{K}{2}} = \sqrt{\frac{|\mathcal{A}|}{2}}.$$

Thus, $\Gamma = \sqrt{\frac{|\mathcal{A}|}{2}}$ is a valid upper bound on the information-ratio coefficient $\Gamma_t$ for every time step $t$. Therefore, by Theorem 2,

$$\mathbb{E}[\mathrm{Reg}_T] \leq \Gamma \sqrt{H(\Pi^\star)\, T} = \sqrt{\frac{|\mathcal{A}|}{2}\, H(\Pi^\star)\, T}. \tag{23}$$

Moreover, since $\Pi^\star$ maps the context space $\mathcal{C}$ to the action space $\mathcal{A}$, the total number of possible mappings is at most $|\mathcal{A}|^{|\mathcal{C}|}$ when $\mathcal{C}$ is finite. This implies that the entropy is bounded as $H(\Pi^\star) \leq \log\big(|\mathcal{A}|^{|\mathcal{C}|}\big) = |\mathcal{C}| \log(|\mathcal{A}|)$. Consequently,

$$\mathbb{E}[\mathrm{Reg}_T] \leq \sqrt{\frac{T|\mathcal{A}||\mathcal{C}| \log |\mathcal{A}|}{2}}.$$

This completes the proof of Corollary 3.

### A.4 Proof of Lemma 4

We fix time step $t$ and the realized context $\mathbf{C}_t = \mathbf{c}_t$. We begin with the definition of the information gain:

$$g_t(a|\mathbf{c}_t) = D_{KL}(P_t(\Pi^*(\mathbf{c}_t), Y_{a,t} \mid \mathbf{c}_t) \,\|\, P_t(\Pi^*(\mathbf{c}_t) \mid \mathbf{c}_t) P_t(Y_{a,t} \mid \mathbf{c}_t)).$$

We compute each term separately.

**Posterior of $\Pi^*(\mathbf{c}_t)$.** By definition,

$$P_t(\Pi^*(\mathbf{c}_t) = a^* \mid \mathbf{c}_t) = P(\Pi^*(\mathbf{c}_t) = a^* \mid \mathcal{F}_t, \mathbf{C}_t = \mathbf{c}_t).$$

Using the law of total expectation over $\boldsymbol{\theta}$,

$$P_t(\Pi^*(\mathbf{c}_t) = a^* \mid \mathbf{c}_t) = \mathbb{E}_{\boldsymbol{\theta} \sim P_t(\boldsymbol{\theta}|\mathbf{c}_t)} \left[ P(\Pi^*(\mathbf{c}_t) = a^* \mid \boldsymbol{\theta}, \mathbf{c}_t) \right].$$

Given $\boldsymbol{\theta}$ and context $\mathbf{c}_t$, the optimal action is deterministic, hence

$$P(\Pi^*(\mathbf{c}_t) = a^* \mid \boldsymbol{\theta}, \mathbf{c}_t) = \mathbb{1}\left\{ a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}[Y_{a,t} \mid \boldsymbol{\theta}, \mathbf{c}_t] \right\}.$$

Substituting,

$$P_t(\Pi^*(\mathbf{c}_t) = a^* \mid \mathbf{c}_t) = \mathbb{E}_{\boldsymbol{\theta} \sim P_t(\boldsymbol{\theta}|\mathbf{c}_t)} \left[ \mathbb{1}\left\{ a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}[Y_{a,t} \mid \boldsymbol{\theta}, \mathbf{c}_t] \right\} \right].$$

**Posterior of $Y_a$.** By definition,

$$P_t(Y_{a,t} = y \mid \mathbf{c}_t) = P(Y_{a,t} = y \mid \mathcal{F}_t, \mathbf{C}_t = \mathbf{c}_t).$$

Conditioning on $\boldsymbol{\theta}$ gives

$$P_t(Y_{a,t} = y \mid \mathbf{c}_t) = \mathbb{E}_{\boldsymbol{\theta} \sim P_t(\boldsymbol{\theta}|\mathbf{c}_t)} \left[ P(Y_{a,t} = y \mid \boldsymbol{\theta}, \mathbf{c}_t) \right].$$

**Joint posterior.** Using the law of total expectation,

$$P_t(\Pi^*(\mathbf{c}_t) = a^*, Y_{a,t} = y \mid \mathbf{c}_t) = \mathbb{E}_{\boldsymbol{\theta} \sim P_t(\boldsymbol{\theta}|\mathbf{c}_t)} \Big[ P(\Pi^*(\mathbf{c}_t) = a^*, Y_{a,t} = y \mid \boldsymbol{\theta}, \mathbf{c}_t) \Big].$$

Since $\Pi^*(\mathbf{c}_t)$ is deterministic given $\boldsymbol{\theta}$ and $\mathbf{c}_t$, we have

$$\begin{aligned} P_t(\Pi^*(\mathbf{c}_t) = a^*, Y_{a,t} = y \mid \mathbf{c}_t) = \mathbb{E}_{\boldsymbol{\theta} \sim P_t(\boldsymbol{\theta}|\mathbf{c}_t)} \Big[ & P(\Pi^*(\mathbf{c}_t) = a^* \mid \boldsymbol{\theta}, \mathbf{c}_t) \\ & \times P(Y_{a,t} = y \mid \boldsymbol{\theta}, \mathbf{c}_t) \Big]. \end{aligned}$$

Substituting the indicator representation,

$$\begin{aligned} P_t(\Pi^*(\mathbf{c}_t) = a^*, Y_{a,t} = y \mid \mathbf{c}_t) = \mathbb{E}_{\boldsymbol{\theta} \sim P_t(\boldsymbol{\theta}|\mathbf{c}_t)} \Big[ & \mathbb{1}\{a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}[Y_{a,t} \mid \boldsymbol{\theta}, \mathbf{c}_t]\} \\ & \times P(Y_{a,t} = y \mid \boldsymbol{\theta}, \mathbf{c}_t) \Big]. \end{aligned}$$

This establishes the expressions used in the KL divergence.

**Expected instantaneous regret.** By definition,

$$\Delta_t(a|\mathbf{c}_t) = \mathbb{E}\big[Y_{\Pi^*(\mathbf{c}_t),t} - Y_{a,t} \mid \mathbf{C}_t = \mathbf{c}_t, \mathcal{F}_t\big] .$$

Using the law of total expectation over $\boldsymbol{\theta}$,

$$\Delta_t(a|\mathbf{c}_t) = \mathbb{E}_{\boldsymbol{\theta}\sim P_t(\boldsymbol{\theta}|\mathbf{c}_t)} \left[\mathbb{E}\big[Y_{\Pi^*(\mathbf{c}_t),t} - Y_{a,t} \mid \boldsymbol{\theta}, \mathbf{c}_t\big]\right] .$$

Expanding the first term gives

$$\mathbb{E}[Y_{\Pi^*(\mathbf{c}_t),t} \mid \boldsymbol{\theta}, \mathbf{c}_t] = \sum_{a^*\in\mathcal{A}} \mathbb{1}\left\{a^* = \arg\max_{a\in\mathcal{A}} \mathbb{E}[Y_{a,t} \mid \boldsymbol{\theta}, \mathbf{c}_t]\right\} \times P(Y_{a^*,t} = 1 \mid \boldsymbol{\theta}, \mathbf{c}_t).$$

Also, since the rewards are Bernoulli,

$$\mathbb{E}[Y_{a,t} \mid \boldsymbol{\theta}, \mathbf{c}_t] = P(Y_{a,t} = 1 \mid \boldsymbol{\theta}, \mathbf{c}_t).$$

Substituting both expressions,

$$\Delta_t(a|\mathbf{c}_t) = \mathbb{E}_{\boldsymbol{\theta}\sim P_t(\boldsymbol{\theta}|\mathbf{c}_t)}\Bigg[\sum_{a^*\in\mathcal{A}} \mathbb{1}\left\{a^* = \arg\max_{a\in\mathcal{A}} \mathbb{E}[Y_{a,t} \mid \boldsymbol{\theta}, \mathbf{c}_t]\right\} \times P(Y_{a^*,t} = 1 \mid \boldsymbol{\theta}, \mathbf{c}_t) - P(Y_{a,t} = 1 \mid \boldsymbol{\theta}, \mathbf{c}_t)\Bigg].$$

This completes the proof of Lemma 4.

### A.5 Proof of Lemma 5

Fix time step $t$ and the realized context $\mathbf{C}_t = \mathbf{c}_t$. From Lemma 4, define

$$f(\boldsymbol{\theta}) := \sum_{a^*\in\mathcal{A}} \mathbb{1}\left\{a^* = \arg\max_{a\in\mathcal{A}} \mathbb{E}[Y_{a,t} \mid \boldsymbol{\theta}, \mathbf{c}_t]\right\} P(Y_{a^*,t} = 1 \mid \boldsymbol{\theta}, \mathbf{c}_t) - P(Y_{a,t} = 1 \mid \boldsymbol{\theta}, \mathbf{c}_t).$$

Then

$$\Delta_t(a \mid \mathbf{c}_t) = \mathbb{E}_{\boldsymbol{\theta}\sim P_t(\boldsymbol{\theta}|\mathbf{c}_t)}[f(\boldsymbol{\theta})] .$$

Let $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ be i.i.d. samples from $P_t(\boldsymbol{\theta} \mid \mathbf{c}_t)$, and define the Monte Carlo estimator

$$\widehat{\Delta}_t(a \mid \mathbf{c}_t) = \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}_i).$$

We now bound the range of $f(\boldsymbol{\theta})$. For any fixed $\boldsymbol{\theta}$, exactly one term in the indicator sum equals 1 because ties are broken deterministically. Hence,

$$\sum_{a^*\in\mathcal{A}} \mathbb{1}\{\cdots\}\, P(Y_{a^*,t} = 1 \mid \boldsymbol{\theta}, \mathbf{c}_t) = P\big(Y_{\Pi^*(\mathbf{c}_t),t} = 1 \mid \boldsymbol{\theta}, \mathbf{c}_t\big) .$$

Because $\Pi^*(\mathbf{c}_t)$ maximizes the expected reward under $\boldsymbol{\theta}$ and $\mathbf{c}_t$,

$$P\big(Y_{\Pi^*(\mathbf{c}_t),t} = 1 \mid \boldsymbol{\theta}, \mathbf{c}_t\big) \geq P(Y_{a,t} = 1 \mid \boldsymbol{\theta}, \mathbf{c}_t).$$

Since both probabilities belong to $[0, 1]$, it follows that

$$0 \leq f(\boldsymbol{\theta}) \leq 1.$$

Define $X_i := f(\boldsymbol{\theta}_i)$. Then $X_1, \ldots, X_N$ are independent and satisfy $0 \leq X_i \leq 1$ almost surely, and

$$\mathbb{E}[X_i] = \Delta_t(a \mid \mathbf{c}_t), \qquad \frac{1}{N}\sum_{i=1}^{N} X_i = \widehat{\Delta}_t(a \mid \mathbf{c}_t).$$

Applying Hoeffding's inequality to the sample mean of bounded variables yields, for any $\epsilon > 0$,

$$\mathbb{P}\Big(\Big|\widehat{\Delta}_t(a \mid \mathbf{c}_t) - \Delta_t(a \mid \mathbf{c}_t)\Big| \geq \epsilon\Big) \leq 2\exp\big(-2N\epsilon^2\big)\,.$$

Setting the right-hand side equal to $\delta$, i.e.,

$$\delta = 2\exp\big(-2N\epsilon^2\big)\,,$$

gives

$$\epsilon = \sqrt{\frac{1}{2N}\log\left(\frac{2}{\delta}\right)}.$$

Therefore, with probability at least $1 - \delta$,

$$\Big|\widehat{\Delta}_t(a \mid \mathbf{c}_t) - \Delta_t(a \mid \mathbf{c}_t)\Big| \leq \sqrt{\frac{1}{2N}\log\left(\frac{2}{\delta}\right)} \leq \sqrt{\frac{2}{N}\log\left(\frac{2}{\delta}\right)}.$$

Hence,

$$\mathbb{P}\left(\Big|\widehat{\Delta}_t(a \mid \mathbf{c}_t) - \Delta_t(a \mid \mathbf{c}_t)\Big| \leq \sqrt{\frac{2}{N}\log\left(\frac{2}{\delta}\right)}\right) \geq 1 - \delta.$$

This completes the proof of Lemma 5.

### A.6 Concentration of the Information Gain

*Proof of Lemma 6.* Fix a time step $t$ and condition on the realized context $\mathbf{C}_t = \mathbf{c}_t$. Throughout this section, ties in the definition of the optimal action are resolved using a fixed deterministic tie-breaking rule. We use $P_t(\boldsymbol{\theta} \mid \mathbf{c}_t)$ as shorthand for $P(\boldsymbol{\theta} \mid \mathcal{F}_t, \mathbf{C}_t = \mathbf{c}_t)$.

For each action $a \in \mathcal{A}$, define

$$\begin{aligned} p_t^a(a^*, y) &:= P_t(\Pi^*(\mathbf{c}_t) = a^*, Y_{a,t} = y \mid \mathbf{c}_t)\,, \\ p_t^\star(a^*) &:= P_t(\Pi^*(\mathbf{c}_t) = a^* \mid \mathbf{c}_t)\,, \\ p_t^a(y) &:= P_t(Y_{a,t} = y \mid \mathbf{c}_t)\,, \end{aligned}$$

for $a^* \in \mathcal{A}$ and $y \in \{0, 1\}$. By Lemma 4, the joint posterior probability satisfies

$$p_t^a(a^*, y) = \mathbb{E}_{\boldsymbol{\theta}\sim P_t(\boldsymbol{\theta}\mid\mathbf{c}_t)}\left[\mathbb{1}\left\{a^* = \arg\max_{a'\in\mathcal{A}} \mathbb{E}\left[Y_{a',t} \mid \boldsymbol{\theta}, \mathbf{c}_t\right]\right\} P(Y_{a,t} = y \mid \boldsymbol{\theta}, \mathbf{c}_t)\right]. \tag{24}$$

Let $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}$ be $N$ independent samples from $P_t(\boldsymbol{\theta} \mid \mathbf{c}_t)$, and define

$$a_i^* := \arg\max_{a'\in\mathcal{A}} \mathbb{E}\left[Y_{a',t} \mid \boldsymbol{\theta}^{(i)}, \mathbf{c}_t\right].$$

The corresponding Monte Carlo estimators are

$$\widehat{p}_t^a(a^*, y) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{a_i^* = a^*\} P\Big(Y_{a,t} = y \mid \boldsymbol{\theta}^{(i)}, \mathbf{c}_t\Big), \tag{25}$$

$$\widehat{p}_t^\star(a^*) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{a_i^* = a^*\}, \tag{26}$$

$$\widehat{p}_t^a(y) := \frac{1}{N}\sum_{i=1}^{N} P\Big(Y_{a,t} = y \mid \boldsymbol{\theta}^{(i)}, \mathbf{c}_t\Big). \tag{27}$$

Define the confidence radii

$$\begin{aligned} \epsilon_1(N,\delta) &:= \sqrt{\frac{1}{2N}\log\Big(\frac{12|\mathcal{A}|^2}{\delta}\Big)}, \\ \epsilon_2(N,\delta) &:= \sqrt{\frac{1}{2N}\log\Big(\frac{6|\mathcal{A}|}{\delta}\Big)}, \\ \epsilon_3(N,\delta) &:= \sqrt{\frac{1}{2N}\log\Big(\frac{12|\mathcal{A}|}{\delta}\Big)}. \end{aligned} \tag{28}$$

**Concentration of the posterior probabilities.** The summands in (25) lie in $[0,1]$. Therefore, for any $\varepsilon > 0$, Hoeffding's inequality gives

$$\mathbb{P}(|\widehat{p}_t^a(a^*, y) - p_t^a(a^*, y)| \geq \varepsilon) \leq 2\exp(-2N\varepsilon^2).$$

Applying a union bound over all $a, a^* \in \mathcal{A}$ and $y \in \{0,1\}$ gives, with probability at least $1 - \delta/3$,

$$|\widehat{p}_t^a(a^*, y) - p_t^a(a^*, y)| \leq \epsilon_1(N,\delta), \qquad \forall a, a^* \in \mathcal{A}, \quad y \in \{0,1\}. \tag{29}$$

Similarly, applying Hoeffding's inequality and a union bound over $a^* \in \mathcal{A}$ gives, with probability at least $1 - \delta/3$,

$$|\widehat{p}_t^\star(a^*) - p_t^\star(a^*)| \leq \epsilon_2(N,\delta), \qquad \forall a^* \in \mathcal{A}. \tag{30}$$

Finally, applying Hoeffding's inequality and a union bound over $a \in \mathcal{A}$ and $y \in \{0,1\}$ gives, with probability at least $1 - \delta/3$,

$$|\widehat{p}_t^a(y) - p_t^a(y)| \leq \epsilon_3(N,\delta), \qquad \forall a \in \mathcal{A}, \quad y \in \{0,1\}. \tag{31}$$

Combining (29), (30), and (31) with a union bound shows that all three concentration events hold simultaneously with probability at least $1 - \delta$.

For notational convenience, define

$$\begin{aligned} L_{1,t}^a(a^*, y) &:= \max\{0, \widehat{p}_t^a(a^*, y) - \epsilon_1\}, \\ U_{1,t}^a(a^*, y) &:= \min\{1, \widehat{p}_t^a(a^*, y) + \epsilon_1\}, \\ L_{2,t}(a^*) &:= \max\{0, \widehat{p}_t^\star(a^*) - \epsilon_2\}, \\ U_{2,t}(a^*) &:= \min\{1, \widehat{p}_t^\star(a^*) + \epsilon_2\}, \\ L_{3,t}^a(y) &:= \max\{0, \widehat{p}_t^a(y) - \epsilon_3\}, \\ U_{3,t}^a(y) &:= \min\{1, \widehat{p}_t^a(y) + \epsilon_3\}, \end{aligned}$$

where $\epsilon_i = \epsilon_i(N,\delta)$ for $i \in \{1,2,3\}$.

**Upper confidence bound.** The contextual information gain can be written as

$$g_t(a \mid \mathbf{c}_t) = D_{\mathrm{KL}}\left(p_t^a \,\|\, p_t^\star p_t^a\right), \tag{32}$$

where

$$\left(p_t^\star p_t^a\right)(a^*, y) = p_t^\star(a^*) p_t^a(y).$$

Using

$$D_{\mathrm{KL}}(P\|Q) \le \chi^2(P\|Q) = \sum_x \frac{P(x)^2}{Q(x)} - 1,$$

together with the concentration bounds above, define

$$\overline{g}_t(a \mid \mathbf{c}_t) := \min\left\{\log 2,\, \max\left\{0,\, \sum_{\substack{a^*\in\mathcal{A}\\ y\in\{0,1\}}} \frac{\left(U_{1,t}^a(a^*, y)\right)^2}{L_{2,t}(a^*) L_{3,t}^a(y)} - 1\right\}\right\}. \tag{33}$$

Whenever $L_{2,t}(a^*)L_{3,t}^a(y) = 0$, the corresponding ratio in (33) is interpreted as $+\infty$. The additional bound by $\log 2$ follows from

$$g_t(a \mid \mathbf{c}_t) = I_t(\Pi^*(\mathbf{c}_t); Y_{a,t} \mid \mathbf{c}_t) \le H_t(Y_{a,t} \mid \mathbf{c}_t) \le \log 2.$$

Consequently, on the joint concentration event,

$$g_t(a \mid \mathbf{c}_t) \le \overline{g}_t(a \mid \mathbf{c}_t), \qquad \forall a \in \mathcal{A}. \tag{34}$$

**Lower confidence bound.** Because the KL divergence is jointly convex in its two arguments, a lower confidence bound is obtained from the following convex program:

$$\begin{aligned}
\underline{g}_t(a \mid \mathbf{c}_t) := \underset{P,Q}{\text{minimize}} \quad & D_{\mathrm{KL}}(P\|Q) \\
\text{subject to} \quad & L_{1,t}^a(a^*, y) \le P(a^*, y) \le U_{1,t}^a(a^*, y), \\
& \qquad\qquad \forall a^* \in \mathcal{A}, \quad y \in \{0,1\}, \\
& L_{2,t}(a^*)L_{3,t}^a(y) \le Q(a^*, y) \le U_{2,t}(a^*)U_{3,t}^a(y), \\
& \qquad\qquad \forall a^* \in \mathcal{A}, \quad y \in \{0,1\}, \\
& \sum_{\substack{a^*\in\mathcal{A}\\ y\in\{0,1\}}} P(a^*, y) = 1, \\
& \sum_{\substack{a^*\in\mathcal{A}\\ y\in\{0,1\}}} Q(a^*, y) = 1, \\
& P(a^*, y) \ge 0, \qquad Q(a^*, y) \ge 0, \\
& \qquad\qquad \forall a^* \in \mathcal{A}, \quad y \in \{0,1\}.
\end{aligned} \tag{35}$$

The KL divergence in (35) is understood in the extended-value sense, with $0\log(0/q) = 0$ and $p\log(p/0) = +\infty$ for $p > 0$.

On the joint concentration event, the true distributions

$$P(a^*, y) = p_t^a(a^*, y)$$

and

$$Q(a^*, y) = p_t^\star(a^*) p_t^a(y)$$

are feasible for (35). Therefore, the optimal value of the program satisfies

$$\underline{g}_t(a \mid \mathbf{c}_t) \le g_t(a \mid \mathbf{c}_t), \qquad \forall a \in \mathcal{A}. \tag{36}$$

Combining (34) and (36), we conclude that, with probability at least $1-\delta$,

$$\underline{g}_t(a \mid \mathbf{c}_t) \le g_t(a \mid \mathbf{c}_t) \le \overline{g}_t(a \mid \mathbf{c}_t), \qquad \forall a \in \mathcal{A}.$$

The numerical solution of (35) is described in Section A.7. □

### A.7 KL Divergence Minimization under Box Constraints

We fix time step $t$, context $\mathbf{C}_t = \mathbf{c}_t$, and an action $a \in \mathcal{A}$. We consider the problem of minimizing the Kullback–Leibler (KL) divergence between two discrete distributions under the box and normalization constraints induced by the high-probability confidence intervals.

Recall that the contextual information gain is defined as

$$g_t(a \mid \mathbf{c}_t) = D_{KL}(P_t(\Pi^*(\mathbf{c}_t), Y_{a,t} \mid \mathbf{c}_t) \parallel P_t(\Pi^*(\mathbf{c}_t) \mid \mathbf{c}_t) P_t(Y_{a,t} \mid \mathbf{c}_t)).$$

To obtain a lower bound on $g_t(a \mid \mathbf{c}_t)$, we solve the following constrained optimization problem over probability distributions $P$ and $Q$:

$$\begin{aligned}
\textbf{minimize} \quad & D_{KL}(P \parallel Q) \\
\textbf{subject to} \quad & L^a_{1,t}(a^*, y) \le P(a^*, y) \le U^a_{1,t}(a^*, y), \\
& \qquad \forall a^* \in \mathcal{A},\ y \in \{0,1\}, \\
& L_{2,t}(a^*) L^a_{3,t}(y) \le Q(a^*, y) \le U_{2,t}(a^*) U^a_{3,t}(y), \\
& \qquad \forall a^* \in \mathcal{A},\ y \in \{0,1\}, \\
& \sum_{a^* \in \mathcal{A},\ y \in \{0,1\}} P(a^*, y) = 1, \qquad \sum_{a^* \in \mathcal{A},\ y \in \{0,1\}} Q(a^*, y) = 1, \\
& P(a^*, y) \ge 0, \quad Q(a^*, y) \ge 0, \quad \forall a^* \in \mathcal{A},\ y \in \{0,1\}.
\end{aligned}$$

Here, the confidence radii are

$$\epsilon_1(N, \delta) = \sqrt{\frac{1}{2N} \log\left(\frac{12|\mathcal{A}|^2}{\delta}\right)},$$

$$\epsilon_2(N, \delta) = \sqrt{\frac{1}{2N} \log\left(\frac{6|\mathcal{A}|}{\delta}\right)},$$

$$\epsilon_3(N, \delta) = \sqrt{\frac{1}{2N} \log\left(\frac{12|\mathcal{A}|}{\delta}\right)}.$$

The KL divergence is jointly convex in $P$ and $Q$, and all the constraints above are linear. Therefore, the optimization problem is a convex program, and every locally optimal solution is globally optimal.

We solve this program using a standard convex optimization solver that supports the relative-entropy function. In particular, the objective is represented as

$$D_{KL}(P \| Q) = \sum_{a^* \in \mathcal{A},\ y \in \{0,1\}} \text{rel_entr}(P(a^*, y), Q(a^*, y)),$$

where

$$\text{rel_entr}(p, q) = p \log\left(\frac{p}{q}\right)$$

is interpreted in the extended-value sense. The box, nonnegativity, and normalization constraints are supplied directly to the solver.

If a feasible initialization is required, one can project the midpoint of each box onto the corresponding bounded probability simplex. For bounds $L_i \le x_i \le U_i$, the Euclidean projection of a vector $z$ onto the bounded simplex has the form

$$x_i = \min\{U_i, \max\{L_i, z_i - \lambda\}\},$$

where $\lambda$ is chosen so that $\sum_i x_i = 1$. The value of $\lambda$ can be found efficiently by bisection. Unlike independently clamping and then renormalizing the coordinates, this projection satisfies the box and normalization constraints simultaneously.

The convex program is solved separately for every $a \in \mathcal{A}$, and its optimal value is used as $\underline{g}_t(a \mid \mathbf{c}_t)$.

### A.8 Proof of Theorem 7 (Contextual Case)

Fix a horizon $T \in \mathbb{N}$ and a realized context sequence

$$\mathbf{C}_{1:T} = (\mathbf{C}_1, \dots, \mathbf{C}_T) = (\mathbf{c}_1, \dots, \mathbf{c}_T).$$

For notational convenience, write

$$\mathbb{E}_{\mathbf{c}_{1:T}}[\cdot] := \mathbb{E}[\,\cdot\,|\mathbf{C}_{1:T} = \mathbf{c}_{1:T}].$$

For every $\pi \in S_{|\mathcal{A}|}$, define

$$\Delta_t(\pi \mid \mathbf{c}_t) := \sum_{a\in\mathcal{A}} \pi(a)\Delta_t(a \mid \mathbf{c}_t), \tag{37}$$

$$g_t(\pi \mid \mathbf{c}_t) := \sum_{a\in\mathcal{A}} \pi(a) g_t(a \mid \mathbf{c}_t), \tag{38}$$

where

$$\Delta_t(a \mid \mathbf{c}_t) := \mathbb{E}_t\left[Y_{\Pi^\star(\mathbf{c}_t),t} - Y_{a,t}\middle|\mathbf{C}_t = \mathbf{c}_t\right], \tag{39}$$

$$g_t(a \mid \mathbf{c}_t) := I_t(\Pi^\star(\mathbf{c}_t); Y_{a,t}|\mathbf{C}_t = \mathbf{c}_t). \tag{40}$$

The corresponding information ratio is

$$\Psi_t(\pi \mid \mathbf{c}_t) := \frac{\Delta_t(\pi \mid \mathbf{c}_t)^2}{g_t(\pi \mid \mathbf{c}_t)}. \tag{41}$$

Let

$$\pi_t^\circ(\mathbf{c}_t) \in \arg\min_{\pi\in S_{|\mathcal{A}|}} \Psi_t(\pi \mid \mathbf{c}_t) \tag{42}$$

denote the oracle IDS distribution, and let $\widehat{\pi}_t(\mathbf{c}_t)$ denote the distribution selected by Algorithm 2 using the confidence sets.

**Step 1: Regret decomposition.** By the tower property,

$$\begin{aligned}&\mathbb{E}_{\mathbf{c}_{1:T}}\left[\sum_{t=1}^{T}\left(Y_{\Pi^\star(\mathbf{C}_t),t} - Y_{A_t,t}\right)\right]\\ &\quad= \mathbb{E}_{\mathbf{c}_{1:T}}\left[\sum_{t=1}^{T}\Delta_t(\widehat{\pi}_t \mid \mathbf{c}_t)\right].\end{aligned} \tag{43}$$

Using

$$\Delta_t(\widehat{\pi}_t \mid \mathbf{c}_t) = \sqrt{\Psi_t(\widehat{\pi}_t \mid \mathbf{c}_t)}\sqrt{g_t(\widehat{\pi}_t \mid \mathbf{c}_t)},$$

followed by Cauchy–Schwarz and Jensen's inequality, gives

$$\mathbb{E}_{\mathbf{c}_{1:T}}\left[\sum_{t=1}^{T}\left(Y_{\Pi^\star(\mathbf{C}_t),t}-Y_{A_t,t}\right)\right] \leq \sqrt{\mathbb{E}_{\mathbf{c}_{1:T}}\left[\sum_{t=1}^{T}\Psi_t(\widehat{\pi}_t \mid \mathbf{c}_t)\right]}\sqrt{\mathbb{E}_{\mathbf{c}_{1:T}}\left[\sum_{t=1}^{T}g_t(\widehat{\pi}_t \mid \mathbf{c}_t)\right]}. \tag{44}$$

**Step 2: Bounding the cumulative information gain.** Let

$$Z_t := \Pi^\star(\mathbf{c}_t).$$

Since $Z_t$ is a deterministic function of the random policy $\Pi^\star$, the data-processing inequality gives

$$I_t(Z_t; Y_{a,t}|\mathbf{C}_t = \mathbf{c}_t) \leq I_t(\Pi^\star; Y_{a,t}|\mathbf{C}_t = \mathbf{c}_t)\,. \tag{45}$$

Consequently,

$$\begin{aligned} g_t(\widehat{\pi}_t \mid \mathbf{c}_t) &= \sum_{a\in\mathcal{A}}\widehat{\pi}_t(a \mid \mathbf{c}_t) I_t(Z_t; Y_{a,t}|\mathbf{C}_t = \mathbf{c}_t) \\ &\leq I_t(\Pi^\star; Y_{A_t,t}|A_t, \mathbf{C}_t = \mathbf{c}_t)\,. \end{aligned} \tag{46}$$

Conditional on the history $\mathcal{F}_t$ and the observed context, the randomization used to generate $A_t$ is independent of $\Pi^\star$. Therefore,

$$I(\Pi^\star; A_t|\mathcal{F}_t, \mathbf{C}_{1:T} = \mathbf{c}_{1:T}) = 0.$$

It follows from the chain rule for mutual information that

$$\begin{aligned} &\mathbb{E}_{\mathbf{c}_{1:T}}\left[\sum_{t=1}^{T} g_t(\widehat{\pi}_t \mid \mathbf{c}_t)\right] \\ &\leq \sum_{t=1}^{T} I(\Pi^\star; A_t, Y_{A_t,t}|\mathcal{F}_t, \mathbf{C}_{1:T} = \mathbf{c}_{1:T}) \\ &= I(\Pi^\star; \mathcal{F}_{T+1}|\mathbf{C}_{1:T} = \mathbf{c}_{1:T}) \\ &\leq H(\Pi^\star|\mathbf{C}_{1:T} = \mathbf{c}_{1:T}) \leq H(\Pi^\star). \end{aligned} \tag{47}$$

Substituting (47) into (44) yields

$$\mathbb{E}_{\mathbf{c}_{1:T}}\left[\sum_{t=1}^{T}\left(Y_{\Pi^\star(\mathbf{C}_t),t}-Y_{A_t,t}\right)\right] \leq \sqrt{H(\Pi^\star)\,\mathbb{E}_{\mathbf{c}_{1:T}}\left[\sum_{t=1}^{T}\Psi_t(\widehat{\pi}_t \mid \mathbf{c}_t)\right]}. \tag{48}$$

**Step 3: Oracle information-ratio bound.** Fix $t$ and let

$$p_t(a \mid \mathbf{c}_t) := P_t(\Pi^\star(\mathbf{c}_t) = a|\mathbf{C}_t = \mathbf{c}_t)\,. \tag{49}$$

Consider the Thompson-sampling distribution

$$\pi_t^{\mathrm{TS}}(a \mid \mathbf{c}_t) = p_t(a \mid \mathbf{c}_t).$$

Because $\pi_t^\circ$ minimizes the true information ratio,

$$\Psi_t(\pi_t^\circ \mid \mathbf{c}_t) \leq \Psi_t(\pi_t^{\mathrm{TS}} \mid \mathbf{c}_t). \tag{50}$$

For $a, b \in \mathcal{A}$, let

$$\mu_{t,a} := \mathbb{E}_t[Y_{a,t} | \mathbf{C}_t = \mathbf{c}_t]\,, \tag{51}$$

$$\mu_{t,a|b} := \mathbb{E}_t[Y_{a,t} | \Pi^\star(\mathbf{c}_t) = b, \mathbf{C}_t = \mathbf{c}_t]\,. \tag{52}$$

Posterior matching gives

$$\Delta_t(\pi_t^{\mathrm{TS}} \mid \mathbf{c}_t) = \sum_{a \in \mathcal{A}} p_t(a \mid \mathbf{c}_t) \left(\mu_{t,a|a} - \mu_{t,a}\right). \tag{53}$$

Hence,

$$\Delta_t(\pi_t^{\mathrm{TS}} \mid \mathbf{c}_t)^2 \le |\mathcal{A}| \sum_{a \in \mathcal{A}} p_t(a \mid \mathbf{c}_t)^2 \left(\mu_{t,a|a} - \mu_{t,a}\right)^2. \tag{54}$$

Since the rewards are Bernoulli, Pinsker's inequality implies

$$\left(\mu_{t,a|a} - \mu_{t,a}\right)^2 \le \frac{1}{2} D_{\mathrm{KL}}(P_t(Y_{a,t} \mid \Pi^\star(\mathbf{c}_t) = a, \mathbf{c}_t) \| P_t(Y_{a,t} \mid \mathbf{c}_t))\,. \tag{55}$$

Therefore,

$$\Delta_t(\pi_t^{\mathrm{TS}} \mid \mathbf{c}_t)^2 \le \frac{|\mathcal{A}|}{2} \sum_{a \in \mathcal{A}} p_t(a \mid \mathbf{c}_t)^2 D_{\mathrm{KL}}(P_t(Y_{a,t} \mid \Pi^\star(\mathbf{c}_t) = a, \mathbf{c}_t) \| P_t(Y_{a,t} \mid \mathbf{c}_t))\,. \tag{56}$$

On the other hand,

$$\begin{aligned} g_t(\pi_t^{\mathrm{TS}} \mid \mathbf{c}_t) &= \sum_{a \in \mathcal{A}} p_t(a \mid \mathbf{c}_t) I_t(\Pi^\star(\mathbf{c}_t); Y_{a,t} | \mathbf{C}_t = \mathbf{c}_t) \\ &= \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} p_t(a \mid \mathbf{c}_t) p_t(b \mid \mathbf{c}_t) \\ &\qquad \times D_{\mathrm{KL}}(P_t(Y_{a,t} \mid \Pi^\star(\mathbf{c}_t) = b, \mathbf{c}_t) \| P_t(Y_{a,t} \mid \mathbf{c}_t))\,. \end{aligned} \tag{57}$$

Every term in (57) is nonnegative. Keeping only the diagonal terms $b = a$ gives

$$g_t(\pi_t^{\mathrm{TS}} \mid \mathbf{c}_t) \ge \sum_{a \in \mathcal{A}} p_t(a \mid \mathbf{c}_t)^2 D_{\mathrm{KL}}(P_t(Y_{a,t} \mid \Pi^\star(\mathbf{c}_t) = a, \mathbf{c}_t) \| P_t(Y_{a,t} \mid \mathbf{c}_t))\,. \tag{58}$$

Combining (56) and (58) yields

$$\Delta_t(\pi_t^{\mathrm{TS}} \mid \mathbf{c}_t)^2 \le \frac{|\mathcal{A}|}{2} g_t(\pi_t^{\mathrm{TS}} \mid \mathbf{c}_t).$$

Thus,

$$\Psi_t(\pi_t^\circ \mid \mathbf{c}_t) \le \Psi_t(\pi_t^{\mathrm{TS}} \mid \mathbf{c}_t) \le \frac{|\mathcal{A}|}{2}. \tag{59}$$

**Step 4: Confidence sets and Algorithm 2.** For vectors $\boldsymbol{\Delta} \in \mathbb{R}^{|\mathcal{A}|}$ and $\boldsymbol{g} \in \mathbb{R}_+^{|\mathcal{A}|}$, define

$$F_t(\pi; \boldsymbol{\Delta}, \boldsymbol{g}) := \frac{(\pi^\top \boldsymbol{\Delta})^2}{\pi^\top \boldsymbol{g}}. \tag{60}$$

At time $t$, Algorithm 2 selects vectors

$$\widetilde{\boldsymbol{\Delta}}_t(\mathbf{c}_t) \in \mathcal{C}_t^{\Delta}(\mathbf{c}_t) \qquad \text{and} \qquad \widetilde{\boldsymbol{g}}_t(\mathbf{c}_t) \in \mathcal{C}_t^{g}(\mathbf{c}_t)$$

by selecting one value from each corresponding confidence interval. Define the information ratio computed from these selected values as

$$\widehat{\Psi}_t(\pi \mid \mathbf{c}_t) := F_t\Big(\pi; \widetilde{\boldsymbol{\Delta}}_t(\mathbf{c}_t), \widetilde{\boldsymbol{g}}_t(\mathbf{c}_t)\Big)\,. \tag{61}$$

Algorithm 2 then selects

$$\widehat{\pi}_t(\mathbf{c}_t) \in \arg\min_{\pi \in S_{|\mathcal{A}|}} \widehat{\Psi}_t(\pi \mid \mathbf{c}_t). \tag{62}$$

Let $\mathcal{E}$ be the event on which, simultaneously for every $t \le T$ and every $a \in \mathcal{A}$,

$$\Delta_t(a \mid \mathbf{c}_t) \in \mathcal{C}^{\Delta}_{t,a}(\mathbf{c}_t) \quad \text{and} \quad g_t(a \mid \mathbf{c}_t) \in \mathcal{C}^{g}_{t,a}(\mathbf{c}_t).$$

Using

$$\delta = \frac{\delta'}{2T|\mathcal{A}|}$$

in Lemmas 5 and 6, and applying a union bound over the time steps, actions, and the two concentration results, gives

$$\mathbb{P}(\mathcal{E}) \ge 1 - \delta'. \tag{63}$$

For each time step $t$, define

$$\gamma_t := \sup_{\substack{\pi \in S_{|\mathcal{A}|}, \\ \boldsymbol{\Delta}_t^1(\mathbf{c}_t), \boldsymbol{\Delta}_t^2(\mathbf{c}_t) \in \mathcal{C}_t^{\Delta}(\mathbf{c}_t), \\ \boldsymbol{g}_t^1(\mathbf{c}_t), \boldsymbol{g}_t^2(\mathbf{c}_t) \in \mathcal{C}_t^{g}(\mathbf{c}_t)}} \left| \frac{\left(\pi^\top \boldsymbol{\Delta}_t^1(\mathbf{c}_t)\right)^2}{\pi^\top \boldsymbol{g}_t^1(\mathbf{c}_t)} - \frac{\left(\pi^\top \boldsymbol{\Delta}_t^2(\mathbf{c}_t)\right)^2}{\pi^\top \boldsymbol{g}_t^2(\mathbf{c}_t)} \right|. \tag{64}$$

The ratios in (64) are interpreted using the conventions $0/0 := 0$ and $x/0 := +\infty$ for every $x > 0$.

On $\mathcal{E}$, the true vectors $\boldsymbol{\Delta}_t(\mathbf{c}_t)$ and $\boldsymbol{g}_t(\mathbf{c}_t)$ belong to their respective confidence sets. By construction, the vectors $\widetilde{\boldsymbol{\Delta}}_t(\mathbf{c}_t)$ and $\widetilde{\boldsymbol{g}}_t(\mathbf{c}_t)$ selected by Algorithm 2 also belong to these confidence sets. Consequently, in (64), we may choose

$$\boldsymbol{\Delta}_t^1(\mathbf{c}_t) = \boldsymbol{\Delta}_t(\mathbf{c}_t), \qquad \boldsymbol{g}_t^1(\mathbf{c}_t) = \boldsymbol{g}_t(\mathbf{c}_t),$$

and

$$\boldsymbol{\Delta}_t^2(\mathbf{c}_t) = \widetilde{\boldsymbol{\Delta}}_t(\mathbf{c}_t), \qquad \boldsymbol{g}_t^2(\mathbf{c}_t) = \widetilde{\boldsymbol{g}}_t(\mathbf{c}_t).$$

Therefore,

$$\left| \Psi_t(\pi \mid \mathbf{c}_t) - \widehat{\Psi}_t(\pi \mid \mathbf{c}_t) \right| \le \gamma_t \qquad \text{for every } \pi \in S_{|\mathcal{A}|}. \tag{65}$$

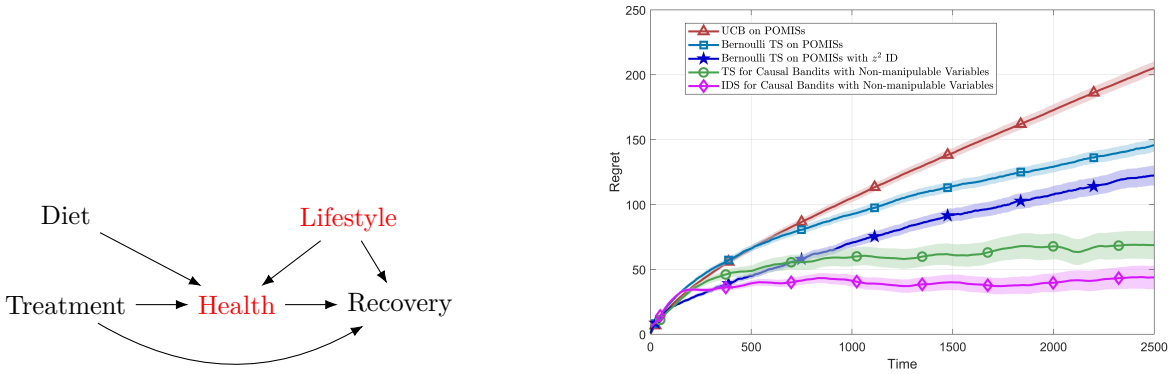
Using (62), (65), and the optimality of $\pi_t^\circ$, we obtain

$$\begin{aligned} \Psi_t(\widehat{\pi}_t \mid \mathbf{c}_t) &\le \widehat{\Psi}_t(\widehat{\pi}_t \mid \mathbf{c}_t) + \gamma_t \\ &\le \widehat{\Psi}_t(\pi_t^\circ \mid \mathbf{c}_t) + \gamma_t \\ &\le \Psi_t(\pi_t^\circ \mid \mathbf{c}_t) + 2\gamma_t \\ &\le \frac{|\mathcal{A}|}{2} + 2\gamma_t. \end{aligned} \tag{66}$$

**Step 5: Final bound.** On the event $\mathcal{E}$, substituting (66) into (48) gives

$$\begin{aligned} &\mathbb{E}\left[ \sum_{t=1}^{T} \left( Y_{\Pi^\star(\mathbf{C}_t),t} - Y_{A_t,t} \right) \,\middle|\, \{\mathbf{C}_t\}_{t=1}^{T} \right] \\ &\qquad \le \sqrt{H(\Pi^\star) \left( \frac{T|\mathcal{A}|}{2} + 2 \sum_{t=1}^{T} \gamma_t \right)}. \end{aligned} \tag{67}$$

Since $\mathbb{P}(\mathcal{E}) \ge 1 - \delta'$, the claimed bound holds with probability at least $1 - \delta'$. This completes the proof.




(a) Causal graph.

(b) Cumulative regret.

Figure 5: Experimental results for a causal graph motivated by a real-world healthcare scenario.

### A.9 Application to a Causal Graph Motivated by a Real-World Scenario

To demonstrate the applicability of our proposed algorithms to practical settings, we consider a causal graph motivated by a real-world healthcare scenario. The graph is illustrated in Figure 5a. The red nodes, Lifestyle and Health, represent non-manipulable variables, while Recovery is the outcome variable of interest. The manipulable variables Diet and Treatment constitute the candidate intervention arms.

We assume that all variables take values in finite discrete domains. We then apply our proposed algorithm, together with the considered baseline methods, to the resulting causal bandit problem. The cumulative regret of the different methods is shown in Figure 5b.

### A.10 Sampling Erdős–Rényi Random Chordal Graphs

To generate connected moral directed acyclic graphs (DAGs), we use a modified Erdős–Rényi sampling procedure. A moral DAG is a DAG whose completed partially directed acyclic graph (CPDAG) consists of a single chain component. Restricting attention to moral DAGs simplifies the orientation procedure while providing a useful setting that can later be extended to more general DAG structures.

We begin by sampling a uniformly random ordering of the vertices, denoted by $\sigma$. For the $n$-th vertex in this ordering, we sample its in-degree according to

$$X_n = \max\{1, \mathrm{Bin}(n-1, \rho)\},$$

where $\mathrm{Bin}(n-1, \rho)$ denotes a binomial random variable with parameters $n-1$ and $\rho$. We then select $X_n$ parents uniformly at random from the vertices that precede the $n$-th vertex in the ordering $\sigma$. The use of at least one parent for every vertex after the first ensures that the resulting graph is connected.

To ensure chordality, we apply the elimination procedure described by Koller & Friedman (2009), using the reverse of $\sigma$ as the elimination ordering. This construction follows the methodology adopted in prior work (Squires et al., 2020). After generating the DAG structure, we independently sample conditional probability tables (CPTs) for all variables in a manner consistent with the parent sets of the sampled graph.