
Causal Discovery for Causal Bandits utilizing Separating Sets

Arnoud A.W.M. de Kroon
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Amsterdam, The Netherlands

Danielle Belgrave
Microsoft Research
Cambridge, United Kingdom

Joris M. Mooij *
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Amsterdam, The Netherlands

Abstract

The Causal Bandit is a variant of the classic Bandit problem where an agent must identify the best action in a sequential decision-making process, where the reward distribution of the actions displays a non-trivial dependence structure that is governed by a causal model. All methods proposed thus far in the literature rely on exact prior knowledge of the causal model to obtain improved estimators for the reward. We formulate a new causal bandit algorithm that is the first to no longer rely on explicit prior causal knowledge and instead uses the output of causal discovery algorithms. This algorithm relies on a new estimator based on separating sets, a causal structure already known in causal discovery literature. We show that given a separating set, this estimator is unbiased, and has lower variance compared to the sample mean. We derive a concentration bound and construct a UCB-type algorithm based on this bound, as well as a Thompson sampling variant. We compare our algorithms with traditional bandit algorithms on simulation data. On these problems, our algorithms show a significant boost in performance.

1 Introduction

In recent years, there have been several works on the Causal Bandit problem (Lattimore et al., 2016; Sen et al., 2017; Yabe et al., 2018; Lee and Bareinboim, 2018). This is a variant of the classical multi-armed bandits problem, where an underlying structural causal model (Pearl, 2009) is assumed between observed variables.

In the bandit problem, we iteratively choose an arm from a set of arms to play, after which we observe a reward variable conditional on the chosen arm. In classical bandits, the rewards for the arms are assumed to be independent. If we assume the rewards are generated by a causal model, the rewards are no longer independent. We can use this additional structure to improve our performance.

Consider the following example. We play a video game with two buttons A and B. The game is played in rounds, and in each round we have to choose which combination of the buttons we push. Then, the game program generates a randomly chosen cute animal S which appears on the screen, for example a giraffe or a zebra, conditional on the buttons pressed. Afterwards, a random cuteness score Y is generated by the program. The distribution of this score is conditional on the animal

*JMM was supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement 639466).

that appeared on the screen. Our goal is then to find out which combinations of buttons to press to maximize the total cuteness score achieved over the course of this game. The generative process of this game can be represented by a causal graph, which is depicted in Figure 1.

This is an example of a causal bandit, where we choose actions and obtain rewards, but in addition to a reward variable, we also observe additional variables after choosing our action. All observed variables are generated through a causal mechanism. If the probability distributions for each button combination of the animals that appear on the screen do not overlap, then the reward signals of the buttons pressed are independent of each other and this game can be considered a classical bandit problem. However, if the distributions overlap, we can share information of the reward signal between these button combinations to better estimate the expected reward value of each combination.

As a concrete example, consider starting the game with no prior knowledge on the conditional distributions. We press button A once, observe a giraffe, and gain 10 cuteness points, and then press button B , again observe a giraffe, but now gain 5 cuteness points. A traditional bandit algorithm would estimate the expected value of button A with the sample mean, which is 10. However, it is intuitively obvious that it is a better strategy to separately model the distribution of the animal that appears on the screen and the expected reward given the animal, thus obtaining an estimated value of button A of 7.5. We will refer to this sharing of data between actions as ‘information leakage’.

Recent approaches to this problem have shown greatly improved regret bounds compared to naïve approaches that treat it like a classical bandit problem by leveraging information leakage (Lattimore et al., 2016; Sen et al., 2017; Yabe et al., 2018). However, they all rely on perfect prior knowledge of the causal structure. In this work, we formulate a causal bandits algorithm which drops the assumption of prior causal knowledge.

In the example, the screen S plays a core role. Once we know how the buttons influence what appears on the screen, we no longer need to know what buttons were pressed to estimate the expected reward: the screen *separates* the action from the reward. This corresponds with the *separating set* concept known from the causality literature (Spirtes et al., 2000; Magliacane et al., 2018; Rojas-Carulla et al., 2018), which (assuming faithfulness) is defined as a set \mathbf{S} that renders a target variable Y independent of a context variable I when conditioned upon: $I \perp\!\!\!\perp Y \mid \mathbf{S}$, where the context variable encodes which interventions are performed. We formulate a Causal Bandit algorithm based on separating sets, where we separately model how actions (i.e. interventions) influence the separating set \mathbf{S} and the expected reward given \mathbf{S} . This will turn out to yield an unbiased estimator, with improved variance compared to a naïve sample mean estimator, on the condition that \mathbf{S} is a correct separating set.

Formulating the algorithm in terms of separating sets allows us to combine it with any causal discovery algorithm that can estimate separating sets from data, and thereby drop the assumption of prior causal knowledge. We formulate a concentration bound for our estimator, and construct an Upper Confidence Bound algorithm based on this bound. We then show greatly improved cumulative regret performance compared to classical bandit algorithms in simulation studies.

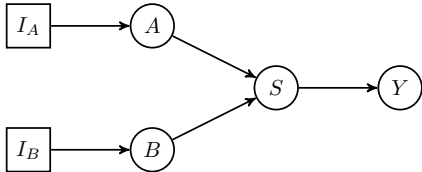


Figure 1: The causal graph for our example game. I_A and I_B are intervention variables encoding interventions on buttons A and B , the screen content is encoded by S , and Y is the reward (cuteness score). $\{S\}$ is a separating set for Y and $\{I_A, I_B\}$, since $\{I_A, I_B\} \perp_G Y \mid S$.

2 Preliminaries

In this section we introduce the required preliminaries regarding causality and causal bandits.

2.1 Causal modeling and graph definitions

We will very briefly introduce the elements of the theory of graphical causal modeling that are used in this work. An in-depth introduction can for example be found in Pearl (2009).

We will denote tuples of variables with a bold capital letter, e.g. $\mathbf{X} = (X_i)_{i=1}^n$, and will use lower case letter x for a value assigned to X . The domain of \mathbf{X} is denoted by $D(\mathbf{X})$. We assume that we observe variables generated through an acyclic Structural Causal Model $\mathcal{M} = \langle \mathbf{V}, \mathbf{E}, \mathbf{F}, \mathbb{P}[\mathbf{E}] \rangle$, with a tuple of endogenous variables \mathbf{V} and a tuple of independent exogenous variables \mathbf{E} with probability distribution $\mathbb{P}[\mathbf{E}]$. The values of \mathbf{V} are defined by the tuple of functions \mathbf{F} , where for each $V_i \in \mathbf{V}$ there is a $f_{V_i} \in \mathbf{F}$ such that $V_i = f_{V_i}(pa(V_i), \mathbf{E}_i)$. Here $pa(V_i) \subseteq \mathbf{V} \setminus \{V_i\}$ are the direct causes (“parents”) of V_i and $\mathbf{E}_i \subseteq \mathbf{E}$ is a subset of the exogenous variables. We explicitly allow for confounders (since the \mathbf{E}_i can overlap), but exclude cycles, though it would be straightforward to include them (see e.g. Mooij et al., 2016). For simplicity, we assume all variables henceforth to be discrete.

Each SCM has an associated graph $G = \langle \mathbf{V}, \mathcal{E} \rangle$, which is acyclic if and only if the SCM is acyclic, where \mathbf{V} is a set of nodes corresponding to the endogenous variables and \mathcal{E} is a set of edges. If V_i directly influences V_j according to f_{V_j} , then there is a directed edge $V_i \rightarrow V_j \in \mathcal{E}$. There is a bidirected edge $V_i \leftrightarrow V_j \in \mathcal{E}$ if they share independent noise variables, i.e., if $\mathbf{E}_i \cap \mathbf{E}_j \neq \emptyset$. We adopt the default family relationships: *pa*, *ch*, *an*, and *de* for parents, children and ancestors and descendants respectively, where for *an* and *de* we include the variable itself.

We may now reason about performing interventions on the variables V_i . In the SCM causal modeling framework, interventions are defined by altering the functional dependencies of the SCM. For example, we may force the value of a variable to a specific value ξ . This is called a *perfect intervention*, and the joint probability is then notated as $\mathbb{P}[\mathbf{V} \mid \text{do}(V_i = \xi)]$. One may also define other types of interventions, for example *soft* interventions which alter the functional dependency f_{V_i} but may keep a functional relationship instead of just setting the variable to a value.

Here we make use of *context variables* (Mooij et al., 2016) to model interventions. We introduce \mathbf{I} to be the set of context variables. We will consider graphs $\mathcal{G} = \langle \mathbf{V} \cup \mathbf{I}, \mathcal{E} \rangle$ with additional vertices \mathbf{I} corresponding to a different interventions. If $I_i \in \mathbf{I}$ encodes an intervention on nodes $\mathbf{T}_i \subseteq \mathbf{V}$, we set I_i to \emptyset if we do not perform this intervention, and to a different value ξ for each possible version of intervention I_i (for example to different perfect intervention values in the domain of \mathbf{T}_i). Furthermore, we add an edge $I_i \rightarrow V_i$ to \mathcal{E} for each $V_i \in \mathbf{T}_i$. For example, we can model a perfect intervention $\text{do}(V_i = \zeta)$ by intervention variable I_i if we modify f_{V_i} to:

$$f_{V_i}^* = \begin{cases} \zeta & \text{if } I_i = \zeta \\ f_{V_i}(pa(V_i), \mathbf{E}_i) & \text{if } I_i = \emptyset \end{cases}$$

Then, if we perform some combination of interventions, this corresponds to choosing a vector of values ζ , of the same size as the number of intervention variables, and where some values may be \emptyset , resulting in $\mathbb{P}[\mathbf{V} \mid \text{do}(\mathbf{I} = \zeta)]$. Note that with this formalism, $\mathbb{P}[\mathbf{V} \mid \text{do}(\mathbf{I} = \zeta)] = \mathbb{P}[\mathbf{V} \mid \mathbf{I} = \zeta]$, because the intervention variables are exogenous.

We define a *path* between nodes V_0 and V_n as a tuple $\langle V_0, e_1, V_1, e_2, \dots, e_n, V_n \rangle$, with $V_i \in \mathbf{V}$, $e_i \in \mathcal{E}$, where each node occurs at most once and e_i is an edge with endpoints V_{i-1} and V_i . V_k is called a *collider* on a path if there is a subpath $\langle V_{k-1}, e_k, V_k, e_{k+1}, V_{k+1} \rangle$ where the edges e_k and e_{k+1} meet head to head on node V_k . Otherwise this node is called a *non-collider*. The endpoints are also referred to as non-colliders.

Using the definition of paths and colliders, one defines d-separation:

Definition 1. (*d-separation*) We say a path $\langle V_0, e_1, \dots, e_n, V_n \rangle$ in graph $\mathcal{G} = \langle \mathbf{V}, \mathcal{E} \rangle$ is blocked by $\mathbf{C} \subseteq \mathbf{V}$ if:

- (i): Its first or last node is in \mathbf{C} , or
- (ii): It contains a collider on a node not in $\text{an}(\mathbf{C})$, or
- (iii): It contains a non-collider in \mathbf{C}

If for sets $\mathbf{A}, \mathbf{B} \subseteq \mathbf{V}$ all paths from nodes in \mathbf{A} to nodes in \mathbf{B} are blocked by $\mathbf{C} \subseteq \mathbf{V}$, we say that \mathbf{A} is d-separated from \mathbf{B} by \mathbf{C} , and write $\mathbf{A} \perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{C}$.

Consider an acyclic SCM \mathcal{M} with graph \mathcal{G} . Let $\mathbb{P}_{\mathcal{M}}$ be the probability distribution induced by this model. Then the Directed Global Markov Property holds for subsets $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{C} \implies \mathbf{A} \perp_{\mathbb{P}_{\mathcal{M}}} \mathbf{B} \mid \mathbf{C}.$$

These conditional independencies are the core information provided by causal reasoning that we exploit in this work. While our algorithm itself does not explicitly assume the converse (called *faithfulness*), this is assumed by many causal discovery algorithms thus we henceforth assume faithfulness as well.

2.2 Causal Bandit problem

The multi-armed bandit problem is one of the classic problems studied in sequential decision making literature (Lai and Robbins, 1985). In this setting, an agent decides on which arm to pull and receives a reward corresponding to that arm. Classically, the rewards of the arms are considered independent which gives rise to strategies like ϵ -greedy, UCB (Auer et al., 2002; Cappé et al., 2013) and Thompson Sampling (Thompson, 1933).

Lattimore, Lattimore, and Reid (2016) introduced the Causal Bandit problem as follows. Consider an agent in a sequential decision making process consisting of T trials. In each trial, the agent chooses an assignment of values ζ to intervention variables \mathbf{I} (also referred to as choosing an arm). It then observes variables from $\mathbb{P}[\mathbf{V} \mid \mathbf{I} = \zeta]$, according to an SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{E}, \mathbf{F}, \mathbb{P}[\mathbf{E}] \rangle$ with corresponding graph $\mathcal{G} = (\mathbf{V} \cup \mathbf{I}, \mathcal{E})$. One of the endogenous variables $Y \in \mathbf{V}$ is the target variable. Thus, when choosing an arm for trial $N + 1$, the agent has observed data $\mathcal{D}^N = \{(\zeta^n, \mathbf{v}^n)\}_{n=1}^N$, which are pairs of intervention node values ζ and realizations of \mathbf{V} . In this paper, we assume all variables to be discrete and for Y to be binary. Let Y^n denote the target variable observed in trial n . The goal is then to minimize the cumulative regret $\mathcal{R} = \sum_{n=1}^T [Y^n - \max_{\zeta} \mathbb{E}[Y \mid \mathbf{I} = \zeta]]$.

As a convenience, we will introduce notation to count the number of samples in our data for which a certain predicate p holds. Let $\mathcal{N}_{\mathcal{D}^N}(p) = |\{(\zeta^n, \mathbf{v}^n) \in \mathcal{D}^N \mid (\zeta^n, \mathbf{v}^n) \models p\}|$. For example, $\mathcal{N}_{\mathcal{D}^N}(Y = 1, \mathbf{I} = \zeta)$ is the number of samples in dataset \mathcal{D}^N for which we performed intervention ζ and observed the value 1 for reward variable Y .

2.3 Related Work

Two types of algorithms have been proposed to solve this problem, those relying on information leakage and those that prune the action space based on the structure of the causal graph. The initial paper by Lattimore et al. (2016) was able to give improved bounds for simple regret for the causal bandit problem compared to traditional methods which assume independent arms. This was done by utilizing *information leakage*: the reward obtained under one intervention may provide information about other interventions. The authors construct an importance sampling estimator based on this principle that assumes full prior knowledge of the probability distribution of all variables besides the target variable. Using this, the authors derive an improved simple regret bound. Sen et al. (2017) focused on applying more advanced techniques from the Bandit literature. For example, they analyze gap dependent bounds and apply dynamic clipping, where they divide the T trials into phases and apply a different clipping constant for each phase. These advances lead to sometimes exponentially better regret than the algorithm by Lattimore et al. (2016).

Yabe et al. (2018) extend Lattimore et al.’s work in a different direction. They consider only binary variables and perfect interventions on subsets of nodes. They use the full knowledge of the graph to estimate the probabilities $p(V \mid pa(V), \mathbf{I} = \zeta)$ for each node $V \in \mathbf{V}$. Interestingly, they only require prior knowledge of the graph and estimate all required probability distributions from data acquired from the actual bandit.

More recently, Lee and Bareinboim (2018) introduced a new method for the causal multi-armed bandit problem. They consider perfect interventions on subsets of nodes of the causal graph. Because they only consider perfect interventions, it is sometimes impossible for some interventions to perform better than other interventions more downstream, and thus they may be pruned.

One thing that all existing approaches have in common is that they assume the causal relationships to be known beforehand, an assumption that is often not met in practice.

3 Separating sets lead to improved estimators

In this section, we generalize the intuition we had about the example game in the introduction to an estimator based on a separating set with favorable properties compared to direct sample mean estimation. We then derive a concentration bound for this estimator.

3.1 The information sharing estimator

The core strategy we have seen in Causal Bandits in previous work is to exploit very specific knowledge about the causal structure in order to construct estimators that share information between arms. In order to make Causal Bandits suitable for causal discovery, we introduce a novel information sharing estimator that relies on less specific knowledge about the causal graph, exploiting information leakage to share data between interventions.

Recall our initial example. The core realization we made is that we may separately estimate the relationship between the combination of buttons pressed and the screen and between the screen and the score. More generally, we say that a set of variables \mathbf{S} is a *separating set* for intervention variables \mathbf{I} and target variable Y if $\mathbf{I} \perp_{\mathcal{G}} Y \mid \mathbf{S}$. By the Markov property and faithfulness, this is equivalent to the conditional independence $\mathbf{I} \perp_{\mathbb{P}_{\mathcal{M}}} Y \mid \mathbf{S}$.

If \mathbf{S} is a separating set, we have for all possible interventions $do(\mathbf{I} = \zeta)$ the following identity by the law of total expectation, where the second equality uses the independence:

$$\begin{aligned} \mathbb{E}[Y \mid \mathbf{I} = \zeta] &= \mathbb{E}[\mathbb{E}[Y \mid \mathbf{S}, \mathbf{I} = \zeta] \mid \mathbf{I} = \zeta] \\ &= \mathbb{E}[\mathbb{E}[Y \mid \mathbf{S}] \mid \mathbf{I} = \zeta]. \end{aligned}$$

We introduce separate estimators $\hat{\mu}(\mathbf{s})$ for $\mathbb{E}[Y \mid \mathbf{S} = \mathbf{s}]$ and $\hat{p}(\mathbf{s} \mid \zeta)$ for $\mathbb{P}[\mathbf{S} = \mathbf{s} \mid \mathbf{I} = \zeta]$. Inspired by the above identity, we then propose the following *information sharing estimator* for $\mathbb{E}[Y \mid \mathbf{I} = \zeta]$:

$$\hat{\mu}_{IS}(\zeta \mid \mathcal{D}^N; \mathbf{S}) := \sum_{\mathbf{s} \in D(\mathbf{S})} \hat{\mu}(\mathbf{s} \mid \mathcal{D}^N) \hat{p}(\mathbf{s} \mid \zeta, \mathcal{D}^N). \quad (1)$$

Since \mathbf{S} is discrete and Y is binary, the sample percentages and mean are the obvious candidates to estimate these quantities. Thus we define:

$$\hat{p}(\mathbf{s} \mid \zeta, \mathcal{D}^N) := \frac{\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}, \mathbf{I} = \zeta)}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)}, \quad (2)$$

$$\hat{\mu}(\mathbf{s} \mid \mathcal{D}^N) := \frac{\mathcal{N}_{\mathcal{D}^N}(Y = 1, \mathbf{S} = \mathbf{s})}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s})}. \quad (3)$$

To further understand the proposed estimator, we estimate its bias and variance. In the appendix we show that the following theorem holds:

Theorem 3.1. *If we calculate $\hat{\mu}_{IS}(\zeta \mid \mathcal{D}^N; \mathbf{S})$ from dataset \mathcal{D}^N as defined above, $\mathbf{I} \perp_{\mathbb{P}_{\mathcal{M}}} Y \mid \mathbf{S}$ and there is at least one sample from each possible intervention, then the information sharing estimator (1) is unbiased and there exists a constant $\alpha^* \in [0, 1)$ such that its variance conditional on the number of samples from each intervention is given by:²*

$$\begin{aligned} \mathbb{V}[\hat{\mu}_{IS}(\zeta \mid \mathcal{D}^N; \mathbf{S})] &= \frac{1}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)} \left(\mathbb{V}_{\mathbf{s} \sim \mathbb{P}[\mathbf{S} = \mathbf{s} \mid \mathbf{I} = \zeta]} [\mathbb{E}[Y \mid \mathbf{S} = \mathbf{s}]] \right. \\ &\quad \left. + (1 - \alpha^*(\zeta, \mathcal{D}^N)) \mathbb{E}_{\mathbf{s} \sim \mathbb{P}[\mathbf{S} = \mathbf{s} \mid \mathbf{I} = \zeta]} [\mathbb{E}[Y \mid \mathbf{S} = \mathbf{s}]] (1 - \mathbb{E}[Y \mid \mathbf{S} = \mathbf{s}]) \right). \end{aligned} \quad (4)$$

Proof. See appendix. □

It is easy to see that if $\alpha^*(\zeta, \mathcal{D}^N) = 0$ (which for example happens if no data with $\mathbf{I} \neq \zeta$ is available) then $\mathbb{V}[\hat{\mu}_{IS}(\zeta \mid \mathcal{D}^N; \mathbf{S})] = \frac{\mathbb{E}[Y \mid \mathbf{I} = \zeta] (1 - \mathbb{E}[Y \mid \mathbf{I} = \zeta])}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)}$, which is the variance of the naïve sample mean calculated only from data where $\mathbf{I} = \zeta$. Thus the information sharing estimator always performs at least as well as the sample mean. We can therefore see the variance of the information sharing

²Here, we abuse notation by omitting explicit conditioning on $\{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)\}$.

estimator in equation (12) as a decomposition of the sample mean variance into first term which can only be reduced by adding more data where $\mathbf{I} = \zeta$, and the second term which can also be reduced by data where $\mathbf{I} \neq \zeta$. Indeed, the first term equals the variance of our estimator if $\hat{\mu}(\mathbf{s}|\mathcal{D}^N)$ would be perfect estimates for the expectations $\mathbb{E}[Y|\mathbf{S} = \mathbf{s}]$.

The second term *can* be reduced by adding data where $\mathbf{I} \neq \zeta$, depending on the overlap in distributions on \mathbf{S} between the interventions. In the appendix we provide lower bounds on $\alpha^*(\zeta, \mathcal{D}^N)$ under different assumptions to better understand when this estimator behaves well, e.g. we show that if we condition on that $\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}) \geq c \cdot \mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}, \mathbf{I} = \zeta)$ for all \mathbf{s} and some positive c , then $\alpha^*(\zeta, \mathcal{D}^N) \geq \frac{c}{1+c}$.

3.2 Concentration bound for information sharing estimator

The true mean $\mu(\zeta) = \mathbb{E}[Y|\mathbf{I} = \zeta]$ is a function of parameters $\mu(\mathbf{s}) = \mathbb{E}[Y|\mathbf{S} = \mathbf{s}]$ and $p(\mathbf{s}|\zeta) = \mathbb{P}[\mathbf{S} = \mathbf{s}|\mathbf{I} = \zeta]$ through the relation $\mu(\zeta) = \sum_{\mathbf{s} \in D(\mathbf{S})} p(\mathbf{s}|\zeta)\mu(\mathbf{s})$. We derive a concentration bound by constraining $p(\mathbf{s}|\zeta)$ and $\mu(\mathbf{s})$ individually with high probability using Hoeffdings bound and the bound on multinomial variables from Weissman et al. (2003). We may then use a union bound on these individual events to obtain a simultaneous multidimensional region Θ of high probability for all parameters. We can then solve the maximization problem:

$$\mathbb{P} \left[\mu(\zeta) \leq \max_{(\mu^*(\mathbf{s}), p^*(\mathbf{s}|\zeta)) \in \Theta} \sum_{\mathbf{s} \in D(\mathbf{S})} p^*(\mathbf{s}|\zeta)\mu^*(\mathbf{s}) \right] \leq \mathbb{P} [(\mu(\mathbf{s}), p(\mathbf{s}|\zeta)) \in \Theta]$$

to obtain a concentration bound. For $\delta \geq 0$, let us define $ucb(\hat{\mu}(\mathbf{s}|\mathcal{D}^N)) = \hat{\mu}(\mathbf{s}|\mathcal{D}^N) + \sqrt{\log(2|D(\mathbf{S})|/\delta)/(2\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}))}$. Moreover, let $\Delta_{\hat{p}}(\zeta) = \sqrt{|S|\log(4/\delta)/(2\mathcal{N}(\mathbf{I} = \zeta))}$. Then the following theorem holds:

Theorem 3.2. *If we calculate $\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S})$ as in (1) from dataset \mathcal{D}^N and $\mathbf{I} \perp_{\mathcal{G}} Y | \mathbf{S}$, then:*

$$\mathbb{P} \left[\mathbb{E}[Y|\mathbf{I} = \zeta] \geq \sum_{\mathbf{s} \in D(\mathbf{S})} \hat{p}(\mathbf{s}|\zeta, \mathcal{D}^N)ucb(\hat{\mu}(\mathbf{s}|\mathcal{D}^N)) + \Delta_{\hat{p}}(\zeta) \left(\max_{\mathbf{s} \in D(\mathbf{S})} ucb(\hat{\mu}(\mathbf{s}|\mathcal{D}^N)) - \min_{\mathbf{s}' \in D(\mathbf{S})} ucb(\hat{\mu}(\mathbf{s}'|\mathcal{D}^N)) \right) \right] \leq \delta. \quad (5)$$

Proof. See appendix. □

4 Separating Set Causal Bandit Algorithms

With a concentration bound in hand, we may now define our Separating Set Causal Bandit UCB algorithm. Note that while the bound of equation (21) is often tighter than the standard bound used in UCB for Bernoulli variables, this is not always the case. Therefore, our algorithm will choose the tightest bound available to it, and will fall back to just the sample mean $\hat{\mu}_{SM}(\zeta|\mathcal{D}^N)$ if the bound of equation (21) is not tighter than the standard UCB bound. To reduce computational cost, we run the causal discovery algorithm once every time the number of iterations has increased by 25%. Each iteration, for each possible intervention, we calculated the normal UCB and for each separating set we calculate the UCB based on equation (21). We then choose as index the UCB which is closest to its corresponding estimate (i.e. it has the smallest width). We then pick the action with the highest index. The full description of the algorithm is in the appendix. Since our UCB algorithm has the same confidence level as used in normal UCB and the bound is at least as tight, we can show the following theorem:

Theorem 4.1. *If we run the separating set causal bandit UCB algorithm as defined in the appendix on dataset \mathcal{D}^N and if $\mathbf{I} \perp_{\mathcal{G}} Y | \mathbf{S}$, it has the same cumulative regret upper-bound as normal UCB.*

Proof. See appendix. □

In practice, as we will see in section 5, this algorithm may perform much better than normal UCB. We also test a Thompson sampling variant of the Separating Set Causal Bandit algorithm. Here,

instead of using an index based on an upper confidence bound, given a separating set, we model the parameters $\mathbb{P}[\mathbf{S} = \mathbf{s} | \mathbf{I} = \zeta]$ using a Dirichlet prior and the parameters $\mathbb{P}[Y = 1 | \mathbf{S} = \mathbf{s}]$ using a Beta prior. We can then apply Thompson sampling, by sampling the parameters from their posterior distributions, and calculating the resulting expected value. Given a sample (ζ, \mathbf{s}, Y) , we can update each of the posteriors separately and naturally. A full specification of this variant is in the appendix.

We may combine our novel algorithms with any causal discovery algorithm which outputs separating sets. The methods we used in our experiments are described in the following subsections.

4.1 Direct Independence Testing

Since we have full interventional data, we can directly test for all sets \mathbf{S} whether they have the separating set property, i.e., whether $\mathbf{I} \perp_{\mathbb{P}_{\mathcal{M}}} Y | \mathbf{S}$. Our baseline causal discovery method is then to directly test for separating sets from data in this way. We here make use of the G^2 -test for conditional independence of discrete variables (Neapolitan, 2004) with p -value threshold $\alpha = 2.5/\sqrt{N}$.

4.2 ASD-JCI123kt

A state-of-the-art causal discovery algorithm for small numbers of variables is ASD-JCI123kt (Mooij et al., 2016). It is a particular implementation of the Joint Causal Inference framework (Mooij et al., 2016), which pools data over contexts. This allows it to simultaneously handle data from different sources, e.g. different interventional distributions. ASD-JCI123kt is a hybrid causal discovery algorithm that scores how well each hypothetical causal graph matches the (strengths of the) observed dependences in the pooled data, giving more weight to stronger dependences. As an independence test, we again make use of the G^2 -test for conditional independence of discrete variables with p -value threshold $\alpha = 2.5/\sqrt{N}$. Contrary to the direct testing baseline, ASD-JCI123kt combines all conditional independence test results in order to score the underlying causal graph(s). Since the algorithm makes use of an Answer Set Program (ASP) building on work by Hyttinen et al. (2014), it is straightforward to query the ASP optimizer for separating sets (e.g., how much evidence is there that variable V_i is independent of V_j given \mathbf{S}), by applying the feature scoring approach proposed by Magliacane et al. (2016). We accept a set \mathbf{S} as a valid separating set if for all $I \in \mathbf{I}$, the confidence score for the independence $I \perp_{\mathbb{P}_{\mathcal{M}}} Y | \mathbf{S}$ output by ASD-JCI123kt is positive.

5 Experimentation

We now proceed to simulate several Causal Bandit problems. For each experiment, initially all algorithms uniformly pick arms 10 to ensure that all statistical tests are well behaved. Both causal discovery algorithms have hyperparameters in the form of a test threshold α , and ASD-JCI123kt furthermore has a score threshold parameter t . We only test one set of these parameters, where we set $\alpha = 2.5/\sqrt{N}$ (which is somewhat reasonable from experience) and $t = 0$, due to the high cost of hyperparameter tuning in this setting. We leave hyperparameter tuning as a further optimization challenge for the future. We compare to UCB and Thompson Sampling baselines, as well as versions of our algorithm with knowledge of a separating set, namely the parents of the target variable.

5.1 Simulation study design

First, we simulate data inspired by our running example game. We generate data as follows. We consider two buttons A and B , with corresponding intervention nodes I_A and I_B . If $I_A = \emptyset$ or $I_B = \emptyset$, we let our younger brother decide whether to press the corresponding button, which he does independently with 50% probability. If we set I_A to 0 we intervene such that button A is not pressed (i.e., $\text{do}(A = 0)$), if we set I_A to 1 we press the corresponding button (i.e., $\text{do}(A = 1)$). Similarly for I_B . Thus there are $3^2 = 9$ possible actions in this bandit. The screen is a binary variable, generated according to $\mathbb{P}[S = 1 | A = a, B = b] = \frac{1+a+b}{4}$. Finally, we generate Y according to $\mathbb{P}[Y = 1 | S = s] = \frac{1+s}{3}$.

Furthermore, we generated all acyclic causal graphs $G = (\mathbf{V}, \mathcal{E})$ over 4 binary variables with no confounders and compare the cumulative regret, with a similar sampling strategy. We allow perfect interventions on all subsets of variables excluding the target variable, and thus there are $3^3 = 27$ possible actions. We only generate graphs where Y has at least 1 parent (otherwise the regret is

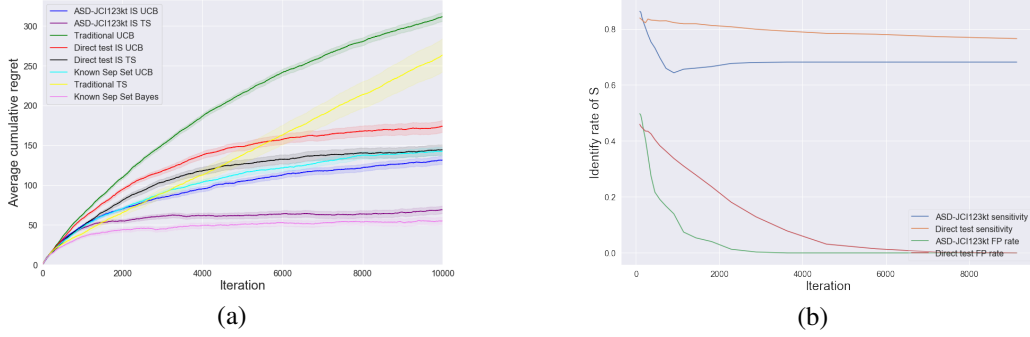


Figure 2: (a): Simulation results on the game example causal bandit over 150 runs. Shaded areas are estimated standard errors. (b): Sensitivity and false positive rate for our causal discovery methods.

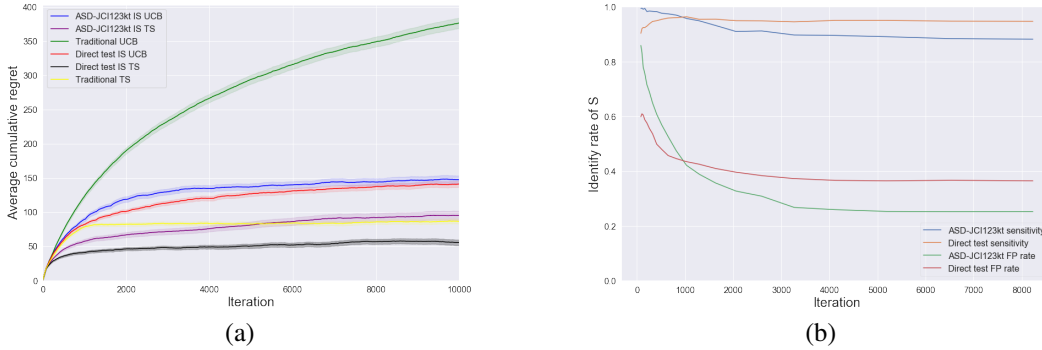


Figure 3: (a): Simulation results on all DAGs of 4 nodes. We generated a parameter sample for each of the 234 graphs. Shaded areas are estimated standard for errors. (b) Sensitivity and false positive rate for our causal discovery methods.

always 0). Permutations of the variables excluding Y are disregarded. Full simulation details are provided in the appendix, including a final simulation study on larger graphs.

5.2 Results

Results for the experiments are shown in figure (2) and (4). As can be seen, traditional UCB is outclassed by all our information sharing (IS) based algorithms that rely on causal discovery. In regimes where the causal discovery methods perform well, Thompson sampling (TS) is also clearly beaten by our IS TS variants and beaten sometimes by our IS UCB variants. The video game example is a structure which seems particularly easy to identify, and therefore the performance of all our IS algorithms is superior on this problem after our causal discovery methods converge.

Unfortunately, in the experiment with all DAGs of 4 nodes, the sensitivity of ASD-JCI123kt converges poorly, likely due to suboptimal hyperparameter settings. Compared to traditional UCB, even with somewhat unreliable causal knowledge our methods show increased performance. However, TS seems to converge quickly for our parameterization strategy after which the mistakes by ASD-JCI123kt are comparatively too costly. Direct testing does converge and therefore our methods using direct testing perform very well, with the TS variant almost immediately converging. We see that surprisingly, even somewhat unreliable causal knowledge may lead to great performance gains, and that this is clearly a very fruitful direction to pursue. However, when the causal discovery methods do not converge properly our estimates are not unbiased and thus in that case there are no convergence guarantees.

6 Conclusion

We have shown that exploiting separating sets in causal bandit problems may yield significantly improved performance compared to traditional UCB and Thompson Sampling. We proved that given correct separating sets, our algorithm has the same regret bound as UCB. In case the causal graph

(and hence, the correct separating sets) is not known, we employed causal discovery algorithms to estimate separating sets from the data in an online fashion. In our simulation experiments, we found that when the causal discovery methods perform reasonably well, our algorithms that rely on them clearly outperform their baseline bandit counterparts.

Our estimator and algorithm may be applied whenever we know of a separating set. This furthermore makes it applicable with pre-existing knowledge less specific than a full causal graph. Furthermore, there is potential in extending this work to contextual bandits and more general reinforcement learning if we formulate an equivalent definition of separating set in these settings. Our approach also turns the Causal Bandit into an interesting task in which to utilize and compare different causal discovery algorithms.

References

- Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.
- Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.
- Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349, 2014.
- Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.
- Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*, pages 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, editors, *Advances in Neural Information Processing Systems 31*, pages 2573–2583. Curran Associates, Inc., 2018.
- Sara Magliacane, Tom Claassen, and Joris M. Mooij. Ancestral causal inference. In *Proceedings of Advances in Neural Information Processing Systems (NIPS 2016)*, pages 4466–4474, Barcelona, Spain, 2016.
- Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10846–10856, 2018.
- Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *arXiv preprint arXiv:1611.10351*, 2016.
- R.E. Neapolitan. *Learning Bayesian Networks*. Pearson, 2004.
- Judea Pearl. *Causality*. Cambridge university press, 2009.
- Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.
- Rajat Sen, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Identifying best interventions through online importance sampling. *arXiv preprint arXiv:1701.02789*, 2017.
- Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.
- 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.
- Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l_1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep.*, 2003.
- 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.

Appendix

A Proof of theorems

In this appendix section, we set out to prove the theorems stated in the main paper. Here it is convenient to introduce vectorized notation for the relevant quantities. Let us consider our estimator given a particular separating set \mathbf{S} with domain $D(\mathbf{S})$. We define the following vectors indexed by $D(\mathbf{S})$, such that the value at index $\mathbf{s} \in \mathbf{S}$ is given by:

$$(\hat{\mathbf{p}}_{\mathbf{S}}(\zeta|\mathcal{D}^N))_{\mathbf{s}} = \hat{p}(\mathbf{s} | \zeta, \mathcal{D}^N), \quad (6)$$

$$(\mathbf{p}_{\mathbf{S}}(\zeta))_{\mathbf{s}} = \mathbb{P}[\mathbf{S} = \mathbf{s} | \mathbf{I} = \zeta], \quad (7)$$

$$(\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N))_{\mathbf{s}} = \hat{\mu}(\mathbf{s} | \mathcal{D}^N), \quad (8)$$

$$(\boldsymbol{\mu}_{\mathbf{S}})_{\mathbf{s}} = \mathbb{E}[Y = 1 | \mathbf{S} = \mathbf{s}], \quad (9)$$

$$(\mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(p))_{\mathbf{s}} = \mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s} \wedge p). \quad (10)$$

With this in hand, we can write the definition of our information sharing estimator (1) as an inner product:

$$\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S}) = \hat{\mathbf{p}}_{\mathbf{S}}^{\top}(\zeta|\mathcal{D}^N) \hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N). \quad (11)$$

A.1 Proof of Theorem 3.1

We set out to prove the theorem:

Theorem 3.1. *If we calculate $\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S})$ from dataset \mathcal{D}^N as defined in (1), $\mathbf{I} \perp_{\mathbb{P}_{\mathcal{M}}} Y | \mathbf{S}$ and there is at least one sample from each possible intervention, then the information sharing estimator (1) is unbiased and there exists a constant $\alpha^* \in [0, 1)$ such that its variance conditional on the number of samples from each intervention is given by:*

$$\begin{aligned} \mathbb{V}[\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S})] &= \frac{1}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)} \left(\mathbb{V}_{\mathbf{s} \sim \mathbb{P}[\mathbf{S}=\mathbf{s} | \mathbf{I}=\zeta]} [\mathbb{E}[Y | \mathbf{S} = \mathbf{s}]] \right. \\ &\quad \left. + (1 - \alpha^*(\zeta, \mathcal{D}^N)) \mathbb{E}_{\mathbf{s} \sim \mathbb{P}[\mathbf{S}=\mathbf{s} | \mathbf{I}=\zeta]} [\mathbb{E}[Y | \mathbf{S} = \mathbf{s}](1 - \mathbb{E}[Y | \mathbf{S} = \mathbf{s}])] \right). \end{aligned} \quad (12)$$

We consider the information sharing estimator (11) calculated from data generated under the random process of the interaction of the policy of a learner with a bandit environment, which we'll denote \mathbb{P}_{ν} , where we assume we have at least one sample from each possible intervention $\zeta \in D(\mathbf{I})$. We first show that the vectors in (11) are uncorrelated, which has as immediate corollary that the information sharing estimator is unbiased. This follows from the law of total expectation:

$$\begin{aligned} \mathbb{E}_{\mathbb{P}_{\nu}} [\hat{\mathbf{p}}_{\mathbf{S}}^{\top}(\zeta|\mathcal{D}^N) \hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N)] &= \mathbb{E}_{\mathbb{P}_{\nu}} \left[\mathbb{E}_{\mathbb{P}_{\nu}} [\hat{\mathbf{p}}_{\mathbf{S}}^{\top}(\zeta|\mathcal{D}^N) \hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N) | \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} = \zeta), \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} \neq \zeta)] \right] \\ &= \mathbb{E}_{\mathbb{P}_{\nu}} \left[\hat{\mathbf{p}}_{\mathbf{S}}^{\top}(\zeta|\mathcal{D}^N) \mathbb{E}_{\mathbb{P}_{\nu}} [\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N) | \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} = \zeta), \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} \neq \zeta)] \right] \\ &= \mathbb{E}_{\mathbb{P}_{\nu}} [\hat{\mathbf{p}}_{\mathbf{S}}^{\top}(\zeta|\mathcal{D}^N)] \boldsymbol{\mu}_{\mathbf{S}} \\ &= \mathbf{p}_{\mathbf{S}}^{\top}(\zeta|\mathcal{D}^N) \boldsymbol{\mu}_{\mathbf{S}} = \mathbb{E}[Y | \mathbf{I} = \zeta] \end{aligned}$$

where in the second line we use that $\hat{\mathbf{p}}_{\mathbf{S}}^{\top}(\zeta|\mathcal{D}^N)$ is deterministic conditional on $\mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} = \zeta)$ and thus it factors out of the inner expectation. On the third line, we use that conditionally on the counts $\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s})$ (which is a deterministic function of the vectors we condition on), $\hat{\mu}(\mathbf{s} | \mathcal{D}^N)$ is just the mean of $\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s})$ Bernoulli variables and thus unbiased, and thus the inner expectation evaluates to the vector of true means $\boldsymbol{\mu}_{\mathbf{S}}(\mathcal{D}^N)$ and factors out. The exact same conditioning argument using the law of total expectation can be used to show that $\mathbb{E}_{\mathbb{P}_{\nu}} [\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N)] = \boldsymbol{\mu}_{\mathbf{S}}$, from which it follows that the expectation factors and thus the vectors are uncorrelated. Finally in the fourth line, the elements of $\hat{\mathbf{p}}_{\mathbf{S}}^{\top}(\zeta|\mathcal{D}^N)$ can be seen as the mean of at least one Bernoulli variable (by assumption) and thus are unbiased, from which the unbiasedness of the information sharing estimator follows.

We analyze the variance using a similar strategy, where we add conditioning through the law of total variance. Let us consider conditioning on the number of samples for each intervention, i.e. we add conditioning on the set $\{\mathcal{N}(\mathbf{I} = \zeta)\}_{\zeta \in D(\mathbf{I})}$:

$$\begin{aligned}\mathbb{V}_{\mathbb{P}_\nu}[\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S})] &= \mathbb{E}_{\mathbb{P}_\nu} \left[\mathbb{V}_{\mathbb{P}_\nu} \left[\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S}) \mid \{\mathcal{N}(\mathbf{I} = \zeta)\}_{\zeta \in D(\mathbf{I})} \right] \right] \\ &\quad + \mathbb{V}_{\mathbb{P}_\nu} \left[\mathbb{E}_{\mathbb{P}_\nu} \left[\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S}) \mid \{\mathcal{N}(\mathbf{I} = \zeta)\}_{\zeta \in D(\mathbf{I})} \right] \right] \\ &= \mathbb{E}_{\mathbb{P}_\nu} \left[\mathbb{V}_{\mathbb{P}_\nu} \left[\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S}) \mid \{\mathcal{N}(\mathbf{I} = \zeta)\}_{\zeta \in D(\mathbf{I})} \right] \right]\end{aligned}$$

where on the first line, the second term is 0 since the estimator is unbiased if we have at least one sample for each possible intervention, which holds by assumption. Thus it suffices to analyze the estimator conditioned on $\{\mathcal{N}(\mathbf{I} = \zeta)\}_{\zeta \in D(\mathbf{I})}$, and then analyze the expectation of the resulting expression w.r.t. \mathbb{P}_ν and the random variables $\{\mathcal{N}(\mathbf{I} = \zeta)\}_{\zeta \in D(\mathbf{I})}$ for a given policy and bandit. We omit explicit conditioning on $\{\mathcal{N}(\mathbf{I} = \zeta)\}_{\zeta \in D(\mathbf{I})}$ to reduce clutter, and analyze the inner variance:

$$\mathbb{V}_{\mathbb{P}_\nu} \left[\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S}) \mid \{\mathcal{N}(\mathbf{I} = \zeta)\}_{\zeta \in D(\mathbf{I})} \right] = \mathbb{V}_{\mathbf{I}} \left[\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S}) \right].$$

Again, we use the law of total variance, adding the same conditioning we did to show unbiasedness:

$$\begin{aligned}\mathbb{V}_{\mathbf{I}} \left[\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S}) \right] &= \mathbb{E}_{\mathbf{I}} \left[\mathbb{V}_{\mathbf{I}} \left[\hat{\mathbf{p}}_{\mathbf{S}}^{\mathbf{I}}(\zeta|\mathcal{D}^N) \hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N) \mid \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} = \zeta), \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} \neq \zeta) \right] \right] \\ &\quad + \mathbb{V}_{\mathbf{I}} \left[\mathbb{E}_{\mathbf{I}} \left[\hat{\mathbf{p}}_{\mathbf{S}}^{\mathbf{I}}(\zeta|\mathcal{D}^N) \hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N) \mid \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} = \zeta), \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} \neq \zeta) \right] \right].\end{aligned}$$

Now note, in both terms, $\hat{\mathbf{p}}_{\mathbf{S}}^{\mathbf{I}}(\zeta|\mathcal{D}^N)$ is non-random. In the second term this vector factors out because of linearity of expectation. In the case of the first term, the individual elements of $\hat{\boldsymbol{\mu}}_{\mathbf{S}}$ are uncorrelated with each-other since they are calculated from disjoint sets of data, thus this vector factors out of the variance element wise squared. This yields:

$$\begin{aligned}\mathbb{V}_{\mathbf{I}} \left[\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S}) \right] &= \mathbb{E}_{\mathbf{I}} \left[\left(\hat{\mathbf{p}}_{\mathbf{S}}^{\mathbf{I}}(\zeta|\mathcal{D}^N) \right)^2 \mathbb{V}_{\mathbf{I}} \left[\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N) \mid \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} = \zeta), \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} \neq \zeta) \right] \right] \quad (13) \\ &\quad + \mathbb{V}_{\mathbf{I}} \left[\hat{\mathbf{p}}_{\mathbf{S}}^{\mathbf{I}}(\zeta|\mathcal{D}^N) \mathbb{E}_{\mathbf{I}} \left[\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N) \mid \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} = \zeta), \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} \neq \zeta) \right] \right]\end{aligned}$$

where the square of the vector in the first term is elementwise, and the variance of a vector in the first term is just the vector of diagonal elements of the covariance matrix, i.e. there are no covariance terms. In the second term, we may now again use that the inner expectation is unbiased following the same argument as before. For the first term, the variance vector $\mathbb{V}_{\mathbf{I}} \left[\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N) \mid \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} = \zeta), \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} \neq \zeta) \right] = \boldsymbol{\mu}_{\mathbf{S}} \otimes (1 - \boldsymbol{\mu}_{\mathbf{S}}) \oslash \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\top)$ is also well defined as the variance of a sample mean of a set of Bernoulli random variables, where \otimes is elementwise product, \oslash is elementwise division and $\mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\top) = \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} = \zeta) + \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} \neq \zeta)$. Substituting this into (13) yields:

$$\begin{aligned}\mathbb{V}_{\mathbf{I}} \left[\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S}) \right] &= \mathbb{E}_{\mathbf{I}} \left[\left(\hat{\mathbf{p}}_{\mathbf{S}}^{\mathbf{I}}(\zeta|\mathcal{D}^N) \right)^2 \oslash \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\top) \right]^{\top} \boldsymbol{\mu}_{\mathbf{S}} \otimes (1 - \boldsymbol{\mu}_{\mathbf{S}}) \quad (14) \\ &\quad + \mathbb{V}_{\mathbf{I}} \left[\hat{\mathbf{p}}_{\mathbf{S}}^{\mathbf{I}}(\zeta|\mathcal{D}^N) \boldsymbol{\mu}_{\mathbf{S}} \right]\end{aligned}$$

Interestingly, the second term corresponds to our information leakage estimator if we were given perfect oracle estimates $\boldsymbol{\mu}_{\mathbf{S}}$. Since the advantage gained by the information sharing estimator is through better estimation of $\boldsymbol{\mu}_{\mathbf{S}}$, this term can be seen as a base error that cannot be reduced through information leakage.

We evaluate the variance of the second term. Let $\mathbf{s}_1, \dots, \mathbf{s}_{\mathcal{N}(\mathbf{I}=\zeta)}$ be the one-hot vector encoded values of \mathbf{S} observed in the subset of our data where $\mathbf{I} = \zeta$. These are i.i.d. categorical variables, and since $\hat{\mathbf{p}}_{\mathbf{S}}^{\mathbf{I}}(\zeta|\mathcal{D}^N) = \frac{1}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I}=\zeta)} \sum_{k=1}^{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I}=\zeta)} \mathbf{s}_k$, the second term becomes by independence:

$$\begin{aligned}\mathbb{V}_{\mathbf{I}} \left[\hat{\mathbf{p}}_{\mathbf{S}}^{\mathbf{I}}(\zeta|\mathcal{D}^N) \boldsymbol{\mu}_{\mathbf{S}} \right] &= \mathbb{V}_{\mathbf{I}} \left[\frac{1}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)} \sum_{k=1}^{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I}=\zeta)} \mathbf{s}_k^{\top} \boldsymbol{\mu}_{\mathbf{S}} \right] \\ &= \frac{1}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)} \mathbb{V}_{\mathbf{I}} \left[\mathbf{s}_1^{\top} \boldsymbol{\mu}_{\mathbf{S}} \right] \\ &= \frac{1}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)} \mathbb{V}_{\mathbf{s} \sim \mathbb{P}[\mathbf{S}=\mathbf{s}|\mathbf{I}=\zeta]} \left[\mathbb{E}[Y|\mathbf{S} = \mathbf{s}] \right]. \quad (15)\end{aligned}$$

Let us now turn our attention to the first term of equation (14). This is an inner product between vectors, where the left factor is an expectation of a vector. Let us consider an element of this expectation vector at index $\mathbf{s} \in D(\mathbf{S})$:

$$\begin{aligned}
\left(\mathbb{E}_{\mathbf{I}} \left[\left(\hat{\mathbf{p}}_{\mathbf{S}}^{\top}(\zeta | \mathcal{D}^N) \right)^2 \oslash \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\top) \right] \right)_{\mathbf{s}} &= \mathbb{E}_{\mathbf{I}} \left[\frac{\hat{p}^2(\mathbf{s} | \zeta, \mathcal{D}^N)}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s})} \right] \\
&= \mathbb{E}_{\mathbf{I}} \left[\frac{\hat{p}(\mathbf{s} | \zeta, \mathcal{D}^N)}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)} \frac{\mathcal{N}(\mathbf{S} = \mathbf{s}, \mathbf{I} = \zeta)}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s})} \right] \\
&= \mathbb{E}_{\mathbf{I}} \left[\frac{\hat{p}(\mathbf{s} | \zeta, \mathcal{D}^N)}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)} \left(1 - \frac{\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}, \mathbf{I} \neq \zeta)}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s})} \right) \right] \\
&= \frac{1}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)} \mathbb{E}_{\mathbf{I}} [\hat{p}(\mathbf{s} | \zeta, \mathcal{D}^N) (1 - \alpha(\mathbf{s}, \zeta, \mathcal{D}^N))] \quad (16)
\end{aligned}$$

where:

$$\alpha(\mathbf{s}, \zeta, \mathcal{D}^N) := \frac{\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}, \mathbf{I} \neq \zeta)}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s})}. \quad (17)$$

Note that $\alpha(\mathbf{s}, \zeta, \mathcal{D}^N)$ equals 0 if $\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}, \mathbf{I} \neq \zeta) = 0$ (i.e. there is no additional data to use where $\mathbf{I} \neq \zeta$ for information sharing for this value of \mathbf{s}), and goes 1 if $\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}, \mathbf{I} \neq \zeta)$ goes to ∞ and we keep $\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}, \mathbf{I} = \zeta)$ fixed, since in the denominator $\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}) = \mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}, \mathbf{I} = \zeta) + \mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}, \mathbf{I} \neq \zeta)$.

Substituting (15) and (16) into (14), the variance of the information sharing estimator then becomes

$$\begin{aligned}
\mathbb{V}_{\mathbf{I}} [\hat{\mu}_{IS}(\zeta | \mathcal{D}^N; \mathbf{S})] &= \frac{1}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)} \left(\mathbb{E}_{\mathbf{I}} [\hat{\mathbf{p}}_{\mathbf{S}}(\zeta, \mathcal{D}^N) \otimes (\mathbf{1} - \alpha(\zeta, \mathcal{D}^N))]^{\top} \boldsymbol{\mu}_S \otimes (\mathbf{1} - \boldsymbol{\mu}_S) \right. \\
&\quad \left. + \mathbb{V}_{\mathbf{s} \sim \mathbb{P}[\mathbf{S}=\mathbf{s} | \mathbf{I}=\zeta]} [\mathbb{E}[Y | \mathbf{S} = \mathbf{s}]] \right) \quad (18)
\end{aligned}$$

where we define $\boldsymbol{\alpha}(\zeta, \mathcal{D}^N)$ as the vectorized version of $\alpha(\mathbf{s}, \zeta, \mathcal{D}^N)$ indexed by $\mathbf{s} \in D(\mathbf{S})$ such that $(\boldsymbol{\alpha}(\zeta, \mathcal{D}^N))_{\mathbf{s}} = \alpha(\mathbf{s}, \zeta, \mathcal{D}^N)$. Let us now consider the term which has $\boldsymbol{\alpha}(\zeta, \mathcal{D}^N)$ as a factor when we expand the parenthesis inside the expectation of the first term. From its definition, we see that $\boldsymbol{\alpha}(\zeta, \mathcal{D}^N)$ is elementwise upper bounded by $\mathbf{1}$ (at an infinite of samples where $\mathbf{I} \neq \zeta$ and a finite number of samples $\mathbf{I} = \zeta$ for all values of \mathbf{s}), and elementwise lower bounded by $\mathbf{0}$ if we have no samples where $\mathbf{I} = \zeta$. Therefore, since all values are positive, if we define:

$$\alpha^*(\zeta, \mathcal{D}^N) = \frac{\mathbb{E}_{\mathbf{I}} [\hat{\mathbf{p}}_{\mathbf{S}}(\zeta, \mathcal{D}^N) \otimes \boldsymbol{\alpha}(\zeta, \mathcal{D}^N)]^{\top} \boldsymbol{\mu}_S \otimes (\mathbf{1} - \boldsymbol{\mu}_S)}{\mathbb{E}_{\mathbf{I}} [\hat{\mathbf{p}}_{\mathbf{S}}(\zeta, \mathcal{D}^N) \otimes \mathbf{1}]^{\top} \boldsymbol{\mu}_S \otimes (\mathbf{1} - \boldsymbol{\mu}_S)} \quad (19)$$

then $\alpha^*(\zeta, \mathcal{D}^N)$ is upper bounded by 1 since from its definition we see that $\boldsymbol{\alpha}(\zeta, \mathcal{D}^N)$ is elementwise upper bounded by $\mathbf{1}$ in which case the numerator and denominator are equal. Furthermore, $\alpha^*(\zeta, \mathcal{D}^N)$ is lower bounded by 0 since all values are nonnegative. Then $\alpha^*(\zeta, \mathcal{D}^N) \in [0, 1]$ and substitution of $\alpha^*(\zeta, \mathcal{D}^N)$ into (18) yields:

$$\begin{aligned}
\mathbb{V}_{\mathbf{I}} [\hat{\mu}_{IS}(\zeta | \mathcal{D}^N; \mathbf{S})] &= \frac{1}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)} \left((1 - \alpha^*(\zeta, \mathcal{D}^N)) \mathbb{E}_{\mathbf{I}} [\hat{\mathbf{p}}_{\mathbf{S}}(\zeta, \mathcal{D}^N) \otimes \mathbf{1}]^{\top} \boldsymbol{\mu}_S \otimes (\mathbf{1} - \boldsymbol{\mu}_S) \right. \\
&\quad \left. + \mathbb{V}_{\mathbf{s} \sim \mathbb{P}[\mathbf{S}=\mathbf{s} | \mathbf{I}=\zeta]} [\mathbb{E}[Y | \mathbf{S} = \mathbf{s}]] \right) \\
&= \frac{1}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)} \left((1 - \alpha^*(\zeta, \mathcal{D}^N)) \mathbf{p}_{\mathbf{S}}^{\top}(\zeta) \boldsymbol{\mu}_S \otimes (\mathbf{1} - \boldsymbol{\mu}_S) + \mathbb{V}_{\mathbf{s} \sim \mathbb{P}[\mathbf{S}=\mathbf{s} | \mathbf{I}=\zeta]} [\mathbb{E}[Y | \mathbf{S} = \mathbf{s}]] \right) \\
&= \frac{1}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)} \left(\mathbb{V}_{\mathbf{s} \sim \mathbb{P}[\mathbf{S}=\mathbf{s} | \mathbf{I}=\zeta]} [\mathbb{E}[Y | \mathbf{S} = \mathbf{s}]] \right. \\
&\quad \left. + (1 - \alpha^*(\zeta, \mathcal{D}^N)) \mathbb{E}_{\mathbf{s} \sim \mathbb{P}[\mathbf{S}=\mathbf{s} | \mathbf{I}=\zeta]} [\mathbb{E}[Y | \mathbf{S} = \mathbf{s}]] (1 - \mathbb{E}[Y | \mathbf{S} = \mathbf{s}]) \right). \quad (20)
\end{aligned}$$

which is what was to be shown. The value of $\alpha^*(\zeta, \mathcal{D}^N)$ is a complicated inner product depending on the model parameters, and is a measure of the expected relative sizes of $\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s} | \mathbf{I} = \zeta)$ and $\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s} | \mathbf{I} \neq \zeta)$ for the values of \mathbf{s} where $\mathbb{P}[\mathbf{S} = \mathbf{s} | \mathbf{I} = \zeta]$ is large.

It is easy to see that $\alpha^*(\zeta, \mathcal{D}^N) \geq \min_{\mathbf{s}} \alpha(\mathbf{s}, \zeta)$, since then $\alpha(\zeta, \mathcal{D}^N) \geq \min_{\mathbf{s}} \alpha(\mathbf{s}, \zeta) \mathbf{1}$ elementwise and we may then factor $\alpha^*(\zeta, \mathcal{D}^N)$ out of the expectation in the numerator of (19) after which the fraction cancels. An interesting case is if we condition on knowing $\{\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}, \mathbf{I} = \zeta)\}_{\mathbf{s} \in D(\mathbf{S})}$. Let us define c to be the largest real number such that for all $\mathbf{s} \in D(\mathbf{S})$, it holds that $\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}, \mathbf{I} \neq \zeta) \geq c \mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s}, \mathbf{I} = \zeta)$. From its definition (17), we see that then $\alpha(\mathbf{s}, \zeta, \mathcal{D}^N) \geq \frac{c}{c+1}$, and thus $\alpha^*(\zeta, \mathcal{D}^N) \geq \frac{c}{c+1}$.

The relative sizes of $\mathbb{V}_{\mathbf{s} \sim \mathbb{P}[\mathbf{S}=\mathbf{s}|\mathbf{I}=\zeta]}[\mathbb{E}[Y|\mathbf{S}=\mathbf{s}]]$ and $\mathbb{E}_{\mathbf{s} \sim \mathbb{P}[\mathbf{S}=\mathbf{s}|\mathbf{I}=\zeta]}[\mathbb{E}[Y|\mathbf{S}=\mathbf{s}](1 - \mathbb{E}[Y|\mathbf{S}=\mathbf{s}])]$ signify how well additional data from $N(\mathbf{I} \neq \zeta)$ helps in estimating $\mathbb{E}[Y|\mathbf{I} = \zeta]$. Interestingly, not always it is even beneficial to share data through information leakage. Specifically, if $\mathbb{E}[Y|\mathbf{S} = \mathbf{s}](1 - \mathbb{E}[Y|\mathbf{S} = \mathbf{s}]) = 0$ for all \mathbf{s} in the support of $\mathbb{P}[\mathbf{S} = \mathbf{s}|\mathbf{I} = \zeta]$, then there is no error due to misestimation of $\mathbb{E}[Y|\mathbf{S} = \mathbf{s}]$ (since they are then deterministic thus if we have just 1 sample this is enough) and all error of the information sharing estimator stems from misestimation of $\mathbb{P}[\mathbf{S} = \mathbf{s}|\mathbf{I} = \zeta]$. In this case, no amount of additional data from $\mathbf{I} \neq \zeta$ may help. In the best case however, the term $\mathbb{V}_{\mathbf{s} \sim \mathbb{P}[\mathbf{S}=\mathbf{s}|\mathbf{I}=\zeta]}[\mathbb{E}[Y|\mathbf{S}=\mathbf{s}]]$ may be 0 in which case all error is reducible through information leakage. This happens for example if $\mathbb{P}[\mathbf{S} = \mathbf{s}|\mathbf{I} = \zeta]$ is deterministic, in which case there is no error due to misestimation of these probabilities.

A.2 Proof of Theorem 3.2

We now set out to show the following theorem:

Theorem 3.2. *If we calculate $\hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S})$ as in (1) from dataset \mathcal{D}^N and $\mathbf{I} \perp_{\mathcal{G}} Y | \mathbf{S}$, then:*

$$\begin{aligned} \mathbb{P} \left[\mathbb{E}[Y|\mathbf{I} = \zeta] > \sum_{\mathbf{s} \in D(\mathbf{S})} \hat{p}(\mathbf{s} | \zeta, \mathcal{D}^N) \text{ucb}(\hat{\mu}(\mathbf{s}|\mathcal{D}^N)) \right. \\ \left. + \Delta_{\hat{p}}(\zeta) \left(\max_{\mathbf{s} \in D(\mathbf{S})} \text{ucb}(\hat{\mu}(\mathbf{s}|\mathcal{D}^N)) - \min_{\mathbf{s}' \in D(\mathbf{S})} \text{ucb}(\hat{\mu}(\mathbf{s}'|\mathcal{D}^N)) \right) \right] < \delta. \end{aligned} \quad (21)$$

Given that \mathbf{S} is a separating set, the true mean $\mu(\zeta) = \mathbb{E}[Y|\mathbf{I} = \zeta]$ is a function of the true parameters $\mathbf{p}_{\mathbf{S}}(\zeta)$ and $\boldsymbol{\mu}_{\mathbf{S}}$:

$$\mu(\zeta) = \mathbf{p}_{\mathbf{S}}(\zeta)^{\top} \boldsymbol{\mu}_{\mathbf{S}}.$$

We will construct an upper bound for $\mu(\zeta)$ by constraining the parameters to some set with high probability, i.e. with high probability $(\mathbf{p}_{\mathbf{S}}(\zeta), \boldsymbol{\mu}_{\mathbf{S}}) \in \Theta = \Theta_{\mathbf{p}} \times \Theta_{\boldsymbol{\mu}}$. Then:

$$\mathbb{P} \left[\mu(\zeta) \leq \sup_{(\mathbf{p}_{\mathbf{S}}^*, \boldsymbol{\mu}_{\mathbf{S}}^*) \in \Theta} (\mathbf{p}_{\mathbf{S}}^*)^{\top} \boldsymbol{\mu}_{\mathbf{S}}^* \right] \geq \mathbb{P}[(\mathbf{p}_{\mathbf{S}}(\zeta), \boldsymbol{\mu}_{\mathbf{S}}) \in \Theta]. \quad (22)$$

We construct Θ by using existing concentration bounds for individual elements of $\mu(\zeta)$'s decomposition. Let us first consider $\boldsymbol{\mu}_{\mathbf{S}}$. We construct an estimator $\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N)$ for these parameters, where each element of this vector is a sample mean of $\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s})$ of Bernoulli variables. Thus, we can use the standard Chernoff-Hoeffding bound for each individual index $\mathbf{s} \in D(\mathbf{S})$ of the vector:

$$\mathbb{P} \left[(\boldsymbol{\mu}_{\mathbf{S}})_{\mathbf{s}} \geq (\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N))_{\mathbf{s}} + \sqrt{\frac{\log(1/\delta_{\mu(\zeta)})}{2\mathcal{N}_{\mathcal{D}^N}(\mathbf{S} = \mathbf{s})}} \right] \leq \delta_{\mu(\zeta)}.$$

Then, we can take the union bound of this event over the indices, to obtain a vectorized complementary version:

$$\mathbb{P} \left[\boldsymbol{\mu}_{\mathbf{S}} < \hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N) + \sqrt{\frac{1}{2} \log(1/\delta_{\mu(\zeta)})} \mathbf{1} \oslash \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\top) \right] \leq 1 - |D(\mathbf{S})| \delta_{\mu(\zeta)},$$

where the square root and the inequality are elementwise, i.e. $\mathbf{a} < \mathbf{b}$ implies that for all \mathbf{s} it holds that $\mathbf{a}_{\mathbf{s}} < \mathbf{b}_{\mathbf{s}}$. We will refer to the complement of the event inside probability as $B_{\boldsymbol{\mu}}$. We now turn to bounding $\mathbf{p}_{\mathbf{S}}(\zeta)$. This is a multinomial variable, i.e. $\hat{\mathbf{p}}_{\mathbf{S}}(\zeta|\mathcal{D}^N) \sim \frac{1}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I}=\zeta)} \text{Multinomial}(\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta), \mathbf{p}_{\mathbf{S}}(\zeta))$. Then by Weissman, Ordentlich, Seroussi, Verdu, and Weinberger (2003):

$$\mathbb{P} \left[\|\hat{\mathbf{p}}_{\mathbf{S}}(\zeta|\mathcal{D}^N) - \mathbf{p}_{\mathbf{S}}(\zeta)\|_1 \geq \sqrt{\frac{2|D(\mathbf{S})| \log(2/\delta_{\mathbf{p}_{\mathbf{S}}})}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta)}} \right] \leq \delta_{\mathbf{p}_{\mathbf{S}}}.$$

We will refer to this event as $B_{\mathbf{p}}$. Then $\mathbb{P}[B_{\boldsymbol{\mu}} \cup B_{\mathbf{p}}] \leq |D(\mathbf{S})|\delta_{\boldsymbol{\mu}(\zeta)} + \delta_{\mathbf{p}_S} := \delta$. It is an interesting optimization problem to choose the values of $\delta_{\boldsymbol{\mu}(\zeta)}$ and $\delta_{\mathbf{p}_S}$ in order to minimize the width of the resulting confidence interval. However, out of convenience we will just pick these such that the confidence is ‘evenly spread out’, i.e. we set $|D(\mathbf{S})|\delta_{\boldsymbol{\mu}(\zeta)} = \delta_{\mathbf{p}_S}$ and set all $\delta_{\boldsymbol{\mu}(\zeta)}$ equal to each other. So then, if we define regions corresponding to the complement of $B_{\boldsymbol{\mu}}$ and $B_{\mathbf{p}}$:

$$\Theta_{\mathbf{p}} = \left\{ \mathbf{p}_S(\zeta) \mid \|\hat{\mathbf{p}}_S(\zeta|\mathcal{D}^N) - \mathbf{p}_S(\zeta)\|_1 < \sqrt{\frac{2|D(\mathbf{S})|\log(4/\delta)}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I}=\zeta)}} \right\},$$

$$\Theta_{\boldsymbol{\mu}} = \left\{ \boldsymbol{\mu}_S \mid \boldsymbol{\mu}_S < \hat{\boldsymbol{\mu}}_S(\mathcal{D}^N) + \sqrt{\frac{1}{2} \log\left(\frac{2|D(\mathbf{S})|}{\delta}\right)} \mathbf{1} \oslash \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\top) \right\}.$$

indeed by union bound:

$$\mathbb{P}[(\mathbf{p}_S(\zeta), \boldsymbol{\mu}_S) \in \Theta] \geq 1 - \delta. \quad (23)$$

It then remains to maximize $\mathbf{p}_S(\zeta)^\top \boldsymbol{\mu}_S$ over $\Theta = \Theta_{\mathbf{p}} \times \Theta_{\boldsymbol{\mu}}$. Let us define $\mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)) = \hat{\boldsymbol{\mu}}_S(\mathcal{D}^N) + \sqrt{\frac{1}{2} \log\left(\frac{2|D(\mathbf{S})|}{\delta}\right)} \mathbf{1} \oslash \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\top)$ and $\Delta_{\hat{p}}(\zeta) = \sqrt{\frac{|D(\mathbf{S})|\log(4/\delta)}{2\mathcal{N}_{\mathcal{D}^N}(\mathbf{I}=\zeta)}}$. We will prove the following lemma:

Lemma A.1. *The optimization problem:*

$$\sup_{(\mathbf{p}_S^*, \boldsymbol{\mu}_S^*) \in \Theta} (\mathbf{p}_S^*)^\top \boldsymbol{\mu}_S^*,$$

under constraints (where inequalities are element-wise):

$$\begin{aligned} \mathbf{p}_S^* &\geq \mathbf{0}, \\ \|\mathbf{p}_S^*\|_1 &= 1, \end{aligned}$$

is upper bounded by:

$$\hat{\mathbf{p}}_S^\top(\zeta|\mathcal{D}^N) \mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)) + \Delta_{\hat{p}}(\zeta) \left(\max_{s \in D(\mathbf{S})} (\mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)))_s - \min_{s' \in D(\mathbf{S})} (\mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)))_{s'} \right)$$

Proof. First note that since all elements of $\mathbf{p}_S^*(\zeta)$ are nonnegative, and we have element-wise upper bounds for $\boldsymbol{\mu}_S^*$, to maximize w.r.t. $\boldsymbol{\mu}_S^*$ we can always just pick the maximum possible value for each element of $\boldsymbol{\mu}_S^*$ in $\Theta_{\boldsymbol{\mu}}$, which are given by $\mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N))$. Thus our maximization problem reduces to:

$$\sup_{(\mathbf{p}_S^*, \boldsymbol{\mu}_S^*) \in \Theta} (\mathbf{p}_S^*)^\top \boldsymbol{\mu}_S^* = \sup_{\mathbf{p}_S^* \in \Theta_{\mathbf{p}}} (\mathbf{p}_S^*)^\top \mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)) \quad (24)$$

Let us define $\Delta_{\mathbf{p}_S} = \mathbf{p}_S^* - \hat{\mathbf{p}}_S^\top(\zeta|\mathcal{D}^N)$. Let Δ_+ be the positive elements of $\Delta_{\mathbf{p}_S}$ and Δ_- be the absolute value of the negative elements of $\Delta_{\mathbf{p}_S}$. Then $\mathbf{p}_S^* = \hat{\mathbf{p}}_S^\top(\zeta|\mathcal{D}^N) + \Delta_+ - \Delta_-$. Substituting this into (24) yields:

$$\sup_{\mathbf{p}_S^* \in \Theta_{\mathbf{p}}} (\mathbf{p}_S^*)^\top \mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)) = \sup_{\mathbf{p}_S^* \in \Theta_{\mathbf{p}}} (\hat{\mathbf{p}}_S^\top(\zeta|\mathcal{D}^N) + \Delta_+ - \Delta_-)^\top \mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)) \quad (25)$$

Now, since $\|\mathbf{p}_S^*\|_1 = 1$ and $\|\hat{\mathbf{p}}_S^\top(\zeta|\mathcal{D}^N)\|_1 = 1$ and all values are positive, it follows that $\|\Delta_+\|_1 = \|\Delta_-\|_1$ and $\|\mathbf{p}_S^* - \hat{\mathbf{p}}_S^\top(\zeta|\mathcal{D}^N)\|_1 = \|\Delta_+\|_1 + \|\Delta_-\|_1$. Looking at the region we are maximizing over, we see that $\|\mathbf{p}_S^* - \hat{\mathbf{p}}_S^\top(\zeta|\mathcal{D}^N)\|_1 \leq \sqrt{\frac{2|D(\mathbf{S})|\log(4/\delta)}{\mathcal{N}_{\mathcal{D}^N}(\mathbf{I}=\zeta)}}$, and thus $\|\Delta_+\|_1 = \|\Delta_-\|_1 \leq \sqrt{\frac{|D(\mathbf{S})|\log(4/\delta)}{2\mathcal{N}_{\mathcal{D}^N}(\mathbf{I}=\zeta)}} = \Delta_{\hat{p}}(\zeta)$. Furthermore, it is trivial to check that for strictly positive values $\Delta_+^\top \mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)) \leq \|\Delta_+\|_1 \max_{s \in D(\mathbf{S})} (\mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)))_s$ and $\Delta_-^\top \mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)) \geq \|\Delta_-\|_1 \min_{s' \in D(\mathbf{S})} (\mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)))_{s'}$. Combining these facts with (25) yields:

$$\begin{aligned} \sup_{\mathbf{p}_S^*(\zeta) \in \Theta_{\mathbf{p}}} \mathbf{p}_S^*(\zeta)^\top \mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)) &\leq \hat{\mathbf{p}}_S^\top(\zeta|\mathcal{D}^N) \mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)) \\ &+ \Delta_{\hat{p}}(\zeta) \left(\max_{s \in D(\mathbf{S})} (\mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)))_s - \min_{s' \in D(\mathbf{S})} (\mathbf{ucb}(\hat{\boldsymbol{\mu}}_S(\mathcal{D}^N)))_{s'} \right) \end{aligned} \quad (26)$$

which finishes the proof. \square

We may now easily combine (22), (23) and (26) to obtain a vectorized version of the theorem statement:

$$\mathbb{P} \left[\mu(\zeta) \leq \hat{\mathbf{p}}_{\mathbf{S}}^{\top}(\zeta | \mathcal{D}^N) \mathbf{ucb}(\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N)) + \Delta_{\hat{p}}(\zeta) \left(\max_{\mathbf{s} \in D(\mathbf{S})} (\mathbf{ucb}(\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N)))_{\mathbf{s}} - \min_{\mathbf{s}' \in D(\mathbf{S})} (\mathbf{ucb}(\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N)))_{\mathbf{s}'} \right) \right] > 1 - \delta \quad (27)$$

which finishes the proof of the theorem. We may follow the exact same argument solving a minimization problem for (22) and taking the reverse Chernoff-Hoeffding bound by defining

$$\mathbf{lcb}(\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N)) = \hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N) - \sqrt{\frac{1}{2} \log \left(\frac{2|D(\mathbf{S})|}{\delta} \right)} \mathbf{1} \oslash \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\top),$$

$$\mathbb{P} \left[\mu(\zeta) \geq \hat{\mathbf{p}}_{\mathbf{S}}^{\top}(\zeta | \mathcal{D}^N) \mathbf{lcb}(\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N)) - \Delta_{\hat{p}}(\zeta) \left(\max_{\mathbf{s} \in D(\mathbf{S})} (\mathbf{lcb}(\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N)))_{\mathbf{s}} - \min_{\mathbf{s}' \in D(\mathbf{S})} (\mathbf{lcb}(\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N)))_{\mathbf{s}'} \right) \right] > 1 - \delta \quad (28)$$

A.3 Proof of theorem 4.1

We set out to prove the theorem:

Theorem 4.1. *If we run the separating set causal bandit UCB algorithm as defined in appendix B.1 on a discrete causal bandit with binary rewards, and if $\mathbf{I} \perp_{\mathcal{G}} Y | \mathbf{S}$, it has the same cumulative regret upper-bound as normal UCB on that bandit.*

To see why this is true, consider the proof of the regret bound for UCB in chapter 8.1 of Lattimore and Szepesvári (2020). To ease notation, this proof is for 1-subgaussian variables, while our problem concerns Bernoulli variables, which are 1/4-subgaussian. This causes an extra factor of 1/4 inside the square of our bounds, but this is an uninteresting technicality. Following that proof and its notation, let μ_i be the expected reward of action i , $\mu^* = \max_i \mu_i = \mu_1$ be the optimal action indexed by 1 for convenience. Then $\Delta_i = \mu^* - \mu_i$ is the expected regret of action i . We consider the regret decomposition at timestep n :

$$R_n = \sum_{i: \Delta_i > 0} \Delta_i \mathbb{E}[T_i(n)]$$

where $T_i(n)$ is the number of times we have chosen action i at timestep n . Let $\hat{\mu}_i(t)$ be the estimated reward of action i (corresponding to some intervention ζ) by the Information Sharing UCB algorithm at timestep t , and $ucb_i(t)$ be the calculated additive upper bound bonus (i.e. the *best_width* calculated in the Information Sharing UCB algorithm). We can then upper-bound $T_i(n)$ as follows:

$$\begin{aligned} T_i(n) &= \sum_{t=1}^n \mathbf{1}\{A_t = i\} \leq \sum_{t=1}^n \mathbf{1}\{\hat{\mu}_1(t-1) + ucb_1(t-1) \leq \mu_1 - \varepsilon\} \\ &\quad + \sum_{t=1}^n \mathbf{1}\{\hat{\mu}_i(t-1) + ucb_i(t-1) > \mu_1 - \varepsilon \text{ and } A_t = i\}, \end{aligned} \quad (29)$$

where A_t is the chosen action, and we have a term corresponding to the number of times the index of the optimal arm is less than $\mu_1 - \varepsilon$ and the second term which corresponds to the number of times that $A_t = i$ and its index is larger than $\mu_1 - \varepsilon$. Let us start with analyzing the expectation of the first term of (29):

$$\mathbb{E} \left[\sum_{t=1}^n \mathbf{1}\{\hat{\mu}_1(t-1) + ucb_1(t-1) \leq \mu_1 - \varepsilon\} \right] \quad (30)$$

Let us assume we picked the information sharing estimator for $\hat{\mu}_1$ at timestep $t-1$. Now, from construction of our algorithm we know that $ucb_1(t) \leq \sqrt{\frac{\log(f(t))}{2T_1(t-1)}}$ where $f(t) = 1 + t \log^2(t)$, where for the information sharing estimator, $ucb_1(t-1)$ is given through the definition of $idx(\zeta, \mathcal{D}^N; \mathbf{S})$ in (43) (after some simplification):

$$ucb_i(t-1) = \hat{\mathbf{p}}^{\top}(\zeta_i, \mathcal{D}^N) \sqrt{\frac{1}{2} \log \left(\frac{2|D(\mathbf{S})|}{\delta} \right)} \mathbf{1} \oslash \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\top) \quad (31)$$

$$+ \sqrt{\frac{|D(\mathbf{S})| \log(4/\delta)}{2\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta_i)}} \left(\max_{\mathbf{s}} (\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N))_{\mathbf{s}} - \min_{\mathbf{s}'} (\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N))_{\mathbf{s}'} \right), \quad (32)$$

where ζ_i is the intervention corresponding to action A_i , and \mathbf{S} is the separating set chosen by the algorithm at timestep $t - 1$ for that intervention. Let us now introduce some further simplifying notation. Let:

$$K_i(t-1) = \hat{\mathbf{p}}^\top(\zeta_i, \mathcal{D}^N) \sqrt{\frac{1}{2} \mathbf{1} \otimes \mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\top)},$$

$$L_i(t-1) = \sqrt{\frac{|D(\mathbf{S})|}{2\mathcal{N}_{\mathcal{D}^N}(\mathbf{I} = \zeta_i)}} \left(\max_{\mathbf{s}} (\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N))_{\mathbf{s}} - \min_{\mathbf{s}'} (\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N))_{\mathbf{s}'} \right)$$

Then:

$$ucb_i(t-1) = K_i(t-1) \sqrt{\log(1/\delta) + \log(2|D(\mathbf{S})|)} + L_i(t-1) \sqrt{\log(1/\delta) + \log(4)} \quad (33)$$

Now, since we picked this upper bound over the sample mean bound, by construction of the algorithm it must hold that at the chosen confidence at parameter at timestep t given by $\delta(t) = 1/f(t)$, it holds that:

$$ucb_i(t-1) = K_i(t-1) \sqrt{\log(f(t)) + \log(2|D(\mathbf{S})|)} + L_i(t-1) \sqrt{\log(f(t)) + \log(4)} \leq \sqrt{\frac{\log(f(t))}{2T_i(t-1)}} \quad (34)$$

Which implies that $K_i(t-1) + L_i(t-1) \leq \sqrt{\frac{1}{2T_i(t-1)}}$. Now consider taking the derivative of $ucb_i(t-1)$ with regard to $\log(1/\delta)$:

$$\frac{\partial}{\partial(\log(1/\delta))} ucb_i(t-1) = \frac{K_i(t-1)}{2\sqrt{\log(1/\delta) + \log(2|D(\mathbf{S})|)}} + \frac{L_i(t-1)}{2\sqrt{\log(1/\delta) + \log(4)}} \quad (35)$$

$$\leq \frac{K_i(t-1) + L_i(t-1)}{2\sqrt{\log(1/\delta)}} \quad (36)$$

$$\leq \frac{1}{2\sqrt{2T_i(t-1)\log(1/\delta)}} = \frac{\partial}{\partial(\log(1/\delta))} \sqrt{\frac{\log(1/\delta)}{2T_i(t-1)}} \quad (37)$$

This shows that the information sharing bound grows more slowly as $\log(1/\delta)$ grows than the traditional UCB bound for $1/4$ -subgaussian variables. Let us now consider $ucb_i(t)$ as a function of δ and calculate δ'_ε such that:

$$\mathbb{P}[\hat{\mu}_i(t-1) + ucb_i(t, 1/(f(t)) + \varepsilon) \leq \mu_1] = \mathbb{P}[\hat{\mu}_i(t-1) + ucb_i(t, \delta'_\varepsilon)] \leq \delta'_\varepsilon, \quad (38)$$

i.e. we solve δ'_ε such that $K_i(t-1) \sqrt{\log(1/\delta'_\varepsilon) + \log(2|D(\mathbf{S})|)} + L_i(t-1) \sqrt{\log(1/\delta'_\varepsilon) + \log(4)} = ucb_i(t, 1/(f(t)) + \varepsilon)$. We may then compare this to analyzing the same event as a $1/4$ -subgaussian variable as in the book, i.e. in that case we solve δ_ε^* such that:

$$\sqrt{\frac{\log(1/\delta_\varepsilon^*)}{2T_i(t-1)}} = \sqrt{\frac{\log(f(t))}{2T_i(t-1)}} + \varepsilon$$

Then, since we have shown that the information sharing bound grows more slowly than the $1/4$ -subgaussian bound as δ decreases, and since clearly $\delta'_\varepsilon \geq 1/(f(t))$ and $\delta_\varepsilon^* \geq 1/(f(t))$, it must hold that $\delta'_\varepsilon \geq \delta_\varepsilon^*$. Thus when we analyze the event:

$$\mathbb{E} \left[\sum_{t=1}^n \mathbf{1} \left\{ \hat{\mu}_1(t-1) + \sqrt{\frac{\log(f(t))}{2T_1(t-1)}} \leq \mu_1 - \varepsilon \right\} \right], \quad (39)$$

where we consider $\hat{\mu}_1(t-1)$ a $1/4$ -subgaussian variable, the resulting upper bound must also be an upper bound for the event $\hat{\mu}_i(t-1) + ucb_i(t, 1/(f(t)) + \varepsilon)$ analyzed with our concentration bound. Substituting this into (39) yields:

$$\mathbb{E} \left[\sum_{t=1}^n \mathbf{1} \{ \hat{\mu}_1(t-1) + ucb_1(t) \leq \mu_1 - \varepsilon \} \right] \leq \mathbb{E} \left[\sum_{t=1}^n \mathbf{1} \left\{ \hat{\mu}_1(t-1) + \sqrt{\frac{\log(f(t))}{2T_1(t-1)}} \leq \mu_1 - \varepsilon \right\} \right] \quad (40)$$

where again very importantly, on the right side we treat $\hat{\mu}_1(t-1)$ as a $1/4$ -subgaussian variable. This analysis was conditional on $\hat{\mu}_1(t-1)$ being an information sharing estimator, however, (40) is trivially true if the algorithm reverted to the sample mean. From here, we may continue the analysis of first term exactly as in the book as the expression is exactly the same modulo the $1/4$ factor inside the square due to the fact that a Bernoulli variable is $1/4$ -subgaussian, while the book focusses on 1 -subgaussian variables to simplify notation. For the analysis of the second term, the following inequality trivially holds:

$$\mathbb{E} \left[\sum_{t=1}^n \mathbf{1} \{ \hat{\mu}_i(t-1) + \text{ucb}_i(t) \geq \mu_1 - \varepsilon \text{ and } A_t = i \} \right] \quad (41)$$

$$\leq \mathbb{E} \left[\sum_{t=1}^n \mathbf{1} \left\{ \hat{\mu}_i(t-1) + \sqrt{\frac{\log(f(t))}{2T_i(t-1)}} \geq \mu_1 - \varepsilon \text{ and } A_t = i \right\} \right], \quad (42)$$

where $\hat{\mu}_i(t-1)$ may be result of the algorithm picking an information sharing estimator or the algorithm reverting to the sample mean. Consider the information sharing estimator. Since its variance is upper bounded by that of the sample mean, and it is bounded between $[0, 1]$, by Hoeffdings lemma it is itself a $1/(4\sqrt{n})$ -subgaussian variable, which implies that corollary 5.5 from the book also holds if $\hat{\mu}_i(t-1)$ is an information sharing estimator. Thus, we may just continue the analysis in the book regardless of the estimator chosen. This concludes the proof of the theorem.

B Algorithm specification

In this section of the appendix, we specify our information sharing Causal Bandit Separating Set algorithms, specifically a UCB and a Thompson sample variant. For both algorithms, we will use the upper bound (27). To ease notation, let us define:

$$\text{idx}(\zeta, \mathcal{D}^N; \mathbf{S}) = \hat{\mathbf{p}}_{\mathbf{S}}^{\top}(\zeta | \mathcal{D}^N) \text{ucb}(\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N)) + \Delta_{\hat{p}}(\zeta) \left(\max_{s \in D(\mathbf{S})} (\text{ucb}(\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N)))_s - \min_{s' \in D(\mathbf{S})} (\text{ucb}(\hat{\boldsymbol{\mu}}_{\mathbf{S}}(\mathcal{D}^N)))_{s'} \right). \quad (43)$$

One important detail when calculating this quantity, is the effect of perfect interventions with known targets. Specifically, consider intervention ζ corresponding to perfect interventions on nodes $\mathbf{O} \subseteq \mathbf{V} \setminus \{Y\}$. If $\mathbf{S} \cap \mathbf{O} = \mathbf{O}'$ is not the empty set, then under intervention ζ , the values for \mathbf{O}' are fixed. This limits the possible support for $\mathbb{P}[\mathbf{S} = \mathbf{s} | \mathbf{I} = \zeta]$, and in the calculation of (43) we may then limit ourselves to values of \mathbf{s} that are possible given the known targets of ζ instead of the full domain $D(\mathbf{S})$, effectively reducing the dimension the problem of estimation of $\mathbb{P}[\mathbf{S} = \mathbf{s} | \mathbf{I} = \zeta]$ by $|\mathbf{O}'|$. This prior knowledge is used when calculating (43) in our experiments and when we construct our Thompson Sampling index.

Both of our algorithms take as input a causal discovery algorithm *disc_sep_set*, that given a dataset \mathcal{D}^N , the set of possible interventions $D(\mathbf{I})$ and the target variable Y attempts to infer the sets \mathbf{S} such that $Y \perp_{\mathcal{G}} \mathbf{I} | \mathbf{S}$ and returns those inferred separating sets.

B.1 Information Sharing UCB

We first define our UCB variant using the information sharing estimator. The full details are in Algorithm 1. First, on line 5, the algorithm retrieves all separating set. Then, for each possible intervention, on line 6 it initializes the best found width so far to the width of the standard naïve UCB bound for Bernoulli variables, and sets the best found index to that of the standard naïve UCB algorithm. For that intervention it then checks each separating set if the bound (27) is tighter, after which on line 17 we store as index for that intervention the index corresponding to the tightest bound. As intervention for that round we then pick the intervention with the highest index.

B.2 Information Sharing TS

Next we define the Thompson Sampling variant of our algorithm. Given a true separating set, it is very natural to construct a Thompson Sampling estimator as follows. We impose a Dirichlet prior on $\mathbf{p}_{\mathbf{S}}(\zeta)$ with parameter $\alpha = 1 \cdot \mathbf{1}$ and for each element of $\boldsymbol{\mu}_{\mathbf{S}}$ we impose a Beta prior with

Algorithm 1 Information Sharing UCB

```
1: Input: Data:  $\mathcal{D}^N = \{(\zeta^n, v^n)\}_{n=1}^N$ , set of possible interventions  $D(\mathbf{I})$ , target variable  $Y$   
   Separating set algorithm:  $disc\_sep\_set$   
2: Output: Next action to pick at iteration  $N + 1$   
3:  $\delta = \frac{1}{1+N \log^2(N)}$   
4: Initialize array  $index[\zeta]$  for  $\zeta \in D(\mathbf{I})$   
5:  $\mathbf{S\_set} \leftarrow disc\_sep\_set(\mathcal{D}^N, Y, D(\mathbf{I}))$   
6: for all  $\zeta \in D(\mathbf{I})$  do  
7:    $best\_width = \sqrt{\frac{\log(1/\delta)}{2N_{\mathcal{D}^N}(\mathbf{I}=\zeta)}}$   
8:    $best\_index = \hat{\mu}_{SM}(\zeta|\mathcal{D}^N) + best\_width$   
9:   for all  $\mathbf{S} \in \mathbf{S\_set}(\zeta)$  do  
10:     $new\_index = idx(\zeta, \mathcal{D}^N; \mathbf{S})$   
11:     $new\_width = new\_index - \hat{\mu}_{IS}(\zeta|\mathcal{D}^N; \mathbf{S})$   
12:    if  $new\_width < best\_width$  then  
13:       $best\_index = new\_index$   
14:       $best\_width = new\_width$   
15:    end if  
16:  end for  
17:   $index[\zeta] = best\_index$   
18: end for  
19: return  $\arg \max_{\zeta} index[\zeta]$ 
```

parameters $\alpha = \beta = 1$, where the parameters are chosen to maximize entropy. We may then calculate the posteriors of these distributions given our data, which are simple and closed form, and then sample from the parameters from them. Given this parameter sample, we may then calculate the corresponding mean and use that as an index.

Unfortunately, our chosen causal discovery methods are inherently frequentist, and thus we did not implement an end-to-end Bayesian approach. Instead, we rely on our Information Sharing UCB algorithm to select a separating set, after which we may construct our Thompson Sampling index assuming that the separating set is correct. If no separating set is found, we revert to a traditional direct Thompson sampling index for that intervention. The full details are given in Algorithm 2, where from line 1-16, we run a variant of Algorithm 1 that just selects for each intervention the separating set that Algorithm 1 would have picked to construct its index. Then, from line 17-20, we construct a Thompson Sampling index for that intervention if a separating set is found.

C Experiments

In this section of the appendix, we specify the details of the experimental design of the experiments with all DAGs of 4 nodes. Furthermore, we detail a final experiment on larger graphs of 6 nodes.

C.1 All DAGs of 4 nodes experiment

When we generate all DAGs of 4 nodes in the manner described in section 5.1, we end up with 234 DAGs. For each generated graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$, we model each node as a binary variable, and add an intervention node I_V for each variable $V \in \mathbf{V} \setminus \{Y\}$. Each intervention node I_V models a perfect intervention on V , where we can either set $I_V = \emptyset$ which corresponds to not intervening on V , we can set it to 1 which corresponds to intervening on V such that $V = 1$, and we can set it to 0 which corresponds to intervening on V such that $V = 0$. For each variable V , we generate a random binary target vector \mathbf{t}_V of size $|pa(V)|$ with uniform distribution. Let $match(\mathbf{t}_V, pa(V))$ be a function that counts the number of parents of V that matches the target vector. We then sample V according to:

$$\mathbb{P}[V = 1|pa(V)] = \frac{1 + match(\mathbf{t}_V, pa(V))}{2 + |pa(V)|}, \quad (44)$$

if we do not intervene on that variable. That is, the probability depends on the numerator which counts the number of parents of V that match the target vector.

Algorithm 2 Information Sharing TS

```

1: Input: Data:  $\mathcal{D}^N = \{(\zeta^n, v^n)\}_{n=1}^N$ , set of possible interventions  $D(\mathbf{I})$ , target variable  $Y$ 
   Separating set algorithm: disc_sep_set
2: Output: Next action to pick at iteration  $N + 1$ 
3:  $\delta = \frac{1}{1+N \log^2(N)}$ 
4: Initialize array index $[\zeta]$  for  $\zeta \in D(\mathbf{I})$ 
5:  $\mathbf{S\_set} \leftarrow \text{disc\_sep\_set}(\mathcal{D}^N, Y, D(\mathbf{I}))$ 
6: for all  $\zeta \in D(\mathbf{I})$  do
7:    $best\_width = \sqrt{\frac{\log(1/\delta)}{2\mathcal{N}_{\mathcal{D}^N}(\mathbf{I}=\zeta)}}$ 
8:    $best\_sep\_set = NULL$ 
9:   for all  $\mathbf{S} \in \mathbf{S\_set}(\zeta)$  do
10:     $new\_index = idx(\zeta, \mathcal{D}^N; \mathbf{S})$ 
11:     $new\_width = new\_index - \hat{\mu}_{IS}(\zeta | \mathcal{D}^N; \mathbf{S})$ 
12:    if  $new\_width < best\_width$  then
13:       $best\_width = new\_width$ 
14:       $best\_sep\_set = \mathbf{S}$ 
15:    end if
16:  end for
17:  if  $best\_sep\_set \neq NULL$  then
18:    Sample  $\tilde{\mathbf{p}} \sim \text{Dirichlet}(\mathbf{N}_{\mathbf{S}, \mathcal{D}^N}(\mathbf{I} = \zeta) + 0.5 \cdot \mathbf{1})$ 
19:    Sample  $\tilde{\boldsymbol{\mu}} \sim (\text{Beta}(\mathcal{N}_{\mathcal{D}^N}(Y = 1, \mathbf{S} = \mathbf{s}) + 1, \mathcal{N}_{\mathcal{D}^N}(Y = 0, \mathbf{S} = \mathbf{s}) + 1))_{\mathbf{s} \in D(\mathbf{S})}$ 
20:     $index[\zeta] = \tilde{\mathbf{p}}^\top \tilde{\boldsymbol{\mu}}$ 
21:  else
22:     $traditional\_TS\_index = \text{Beta}(\mathcal{N}_{\mathcal{D}^N}(Y = 1, \mathbf{I} = \zeta) + 1, \mathcal{N}_{\mathcal{D}^N}(Y = 0, \mathbf{I} = \zeta) + 1)$ 
23:  end if
24: end for
25: return  $\arg \max_{\zeta} index[\zeta]$ 

```

C.2 Experiment with DAGs of 6 nodes

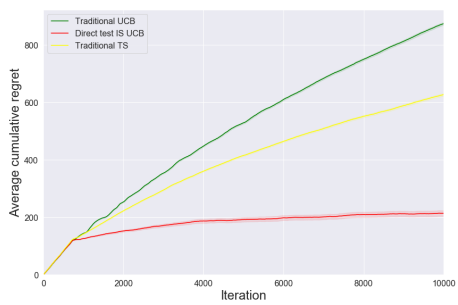
We describe our experiment on randomly generated graphs of 6 nodes. The graph generation process is as follows. We define nodes V_1, \dots, V_6 ordered by their topological ordering, and set $Y = V_6$. For each node V_i with $i > 1$, we randomly sample between one or two from $V_{i' < i}$ with uniform distribution on the nodes selected and the number of nodes sampled.

We allow interventions on all subsets of nodes besides Y , and thus add an intervention variable I_V for all nodes except Y . Each intervention node I_V models a perfect intervention on V , where we can either set $I_V = \emptyset$ which corresponds to not intervening on V , we can set it to 1 which corresponds to intervening on V such that $V = 1$, and we can set it to 0 which corresponds to intervening on V such that $V = 0$. This implies that there are $3^5 = 243$ possible actions. Because of the large action space, we reduced the number of initial pulls for each possible intervention to 3. For each variable V , we generate a random binary target vector \mathbf{t}_V of size $|pa(V)|$ with uniform distribution. Let $match(\mathbf{t}_V, pa(V))$ be a function that counts the number of parents of V that matches the target vector. We then sample V according to:

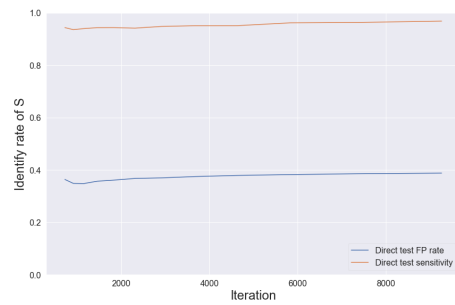
$$\mathbb{P}[V = 1 | pa(V)] = \frac{1 + match(\mathbf{t}_V, pa(V))}{2 + |pa(V)|}, \quad (45)$$

if we do not intervene on that variable. That is, the probability depends on the numerator which counts the number of parents of V that match the target vector.

The results are shown in figure 4. Unfortunately, we did not have enough time to include the Information Sharing Thompson Sampling variant due to long computation time. Furthermore, ASD-JCI123kt is too slow to generalize to this setting. As can be seen, information sharing UCB outperforms baseline methods significantly. The baseline models are slow to converge in this setting due to the high number of actions, while the separating set algorithm performs very well due to the large number of data that can be shared for each action.



(a)



(b)

Figure 4: (a): Simulation results on the game example causal bandit over 150 runs. Shaded areas are estimated standard errors. (b): Sensitivity and false positive rate for our causal discovery methods.