# Transfer Learning in Multi-Armed Bandit: A Causal Approach

Junzhe Zhang
Purdue University
zhang745@purdue.edu

Elias Bareinboim
Purdue University
eb@purdue.edu

## ABSTRACT

Reinforcement learning (RL) agents have been deployed in complex environments where interactions are costly and learning is slow. One prominent task in these settings is to reuse interactions performed by other agents to try to accelerate the learning process. Causal inference provides a family of methods to inferring the effects of actions from a combination of data and qualitative assumptions about the underlying environment. Despite its success of transferring invariant knowledge across domains in the empirical sciences, causal inference has not been fully realized in the context of transfer learning in interactive domains. In this paper, we use causal inference as a basis to support a principled and more robust transfer of knowledge in RL settings. In particular, we tackle the problem of transferring knowledge across bandit agents in settings where causal effects cannot be identified by Pearl's do-calculus and standard learning techniques. Our new identification strategy combines two steps – first, deriving bounds over the arm's distribution based on structural knowledge; second, incorporating these bounds in a dynamic allocation procedure so as to guide the search towards more promising actions. We formally prove that our strategy dominates previously known algorithms and can potentially achieve orders of magnitude faster convergence rates. Finally, we perform simulations and empirically demonstrate that our strategy is consistently more efficient than the current (non-causal) state-of-the-art methods.

## CCS Concepts

**•Computing methodologies → Causal reasoning and diagnostics;**

## Keywords

Causal Inference; Transfer Learning; Reinforcement Learning

## 1. INTRODUCTION

In reinforcement learning (RL), the agent makes a sequence of decisions trying to maximize a particular measure of performance. Typical RL methods train agents in isolation, often taking a substantial amount of time and effort to learn a reasonable control policy. Techniques based on transfer learning (TL) attempt to accelerate the learning process of a target task by reusing knowledge gathered from a different, but somewhat related source task. Common approaches try to exploit various types of domain expertise and transfer knowledge that is invariant across the source and target domains [18, 12, 16]. For a general survey on these techniques, see [15, 28].

Causal inference deals with the problem of inferring the effect of actions (target) from a combination of a causal model (to be defined) and heterogeneous sources of data [22, 6]. One of the fundamental challenges in the field is to determine whether a causal effect can be inferred from the observational (non-experimental) distribution when important variables in the problem may be unmeasured (also called unobserved confounders, or UCs). In fact, qualitative knowledge about cause and effect relationships is often available in complex RL settings. For example, a change of direction of a self-driving car must be caused by a change of the steering wheel, not vice-versa; a surge in users' clicks (click-through rate) causes an observed revenue growth of an advertising engine, not the other way around.

In his seminal work, [21] developed a general calculus known as *do-calculus* by which probabilistic sentences involving interventions and observations can be transformed into other such sentences. The do-calculus was shown to be complete for observational and experimental identification, i.e., any causal effect can be identified from observational or experimental data if and only if it can be reduced to a certain syntactic form in do-calculus [30, 27, 10, 5].

Despite its success in identifying the effect of actions from heterogeneous data in compelling settings across the sciences [6], causal inference techniques have rarely been used to assist the transfer of knowledge in interactive domains. [19, 20] assumed a causal model for the underlying task and performed the transfer of probabilistic knowledge leveraging the invariance encoded in the causal model. Still, they were oblivious to the existence of UCs and had not considered the transfer of causal knowledge. Connections between causal models with UCs and RL were first established in [4]. Nevertheless, these methods mainly dealt with online learning scenarios and barely touched the problem of TL.

In this paper, we marry transfer in RL with the theory of causal inference. We study the offline (batch) transfer problem between two multi-armed bandits (MAB) agents given a causal model of the environment while allowing the existence of UCs. We apply causal inference algorithms to iden-

**Appears in:** *Proceedings of the 1st Workshop on Transfer in Reinforcement Learning (TiRL) at the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), A. Costa, D. Precup, M. Veloso, M. Taylor (chairs), May 8–9, 2017, São Paulo, Brazil.*

tify the causal effect of the target agent's action from trajectories of the source agent. In particular, we study three canonical learning settings where the causal effect is non-identifiable using do-calculus and standard learning techniques, and show that learning speed can still be improved by leveraging prior experiences. Our more detailed contributions are listed below:

1. We formulate the transfer learning across MAB agents in causal language and connect it with the algorithm for identifying causal effects.

2. For three canonical tasks where the causal effect is not identifiable, we provide an efficient method to extract knowledge from the available distributions as bounds over the expected reward (called *causal bounds*).

3. We propose two novel MAB algorithms (B-kl-UCB and B-TS) that take the causal bounds as input. We prove that the regret bound of B-kl-UCB dominates the standard kl-UCB [7]. If the causal bounds impose informative constraints over the arms' distribution, B-kl-UCB will be orders of magnitude faster than kl-UCB; otherwise, the behavior of B-kl-UCB deteriorates to kl-UCB, which we show cannot be improved.

4. We run extensive simulations comparing the proposed strategies (B-kl-UCB and B-TS) against standard MAB solvers and show that our algorithms are consistent and more efficient than state-of-the-art methods.

# 2. PRELIMINARIES

In this section, we introduce the basic notations and definitions used throughout the paper. For a variable $X$, let $D(X)$ denote its domain and $|X|$ denote the dimension of its domain. We will consistently use the abbreviation $P(x)$ for the probabilities $P(X = x), x \in D(X)$.

## 2.1 Multi-Armed Bandits.

The stochastic MAB is a simple yet powerful framework that formalizes the online learning problem with partial feedback [26]. The MAB is described as follows: There are finitely many arms indexed by $a \in \{1, \dots, K\}$ with $K \geq 2$, each associated with an unknown probability distribution $\nu_a$ over $[0, 1]$. Denote the associated expected reward $\mu_a$ of $\nu_a$. At each trial $t = 1 \dots, T$, the agent picks an arm $A_t \in \{1, \dots, K\}$ and receives a reward $Y_t$. Let $a^* = \arg\max_{a \in \{1,\dots,K\}} \mu_a$ and write $\mu^*$ as the expected reward for the optimal arm $a^*$. The number of times each arm $a$ is pulled for $T$ trials is referred as $N_a(T) = \sum_{t=1}^{T} \mathbb{I}\{A_t = a\}$, where $\mathbb{I}\{\cdot\}$ is an indicator function. The quality of a strategy is evaluated through the notion of cumulative regret:

$$R_T = \mathbb{E}[T\mu^* - \sum_{t=1}^{T} Y_T] = T\mu^* - \sum_{t=1}^{T} \mathbb{E}[Y_t] = \sum_{a=1}^{K} \Delta_a \mathbb{E}[N_a(T)]$$

where $\Delta_a = \mu^* - \mu_a$. We use $KL(\mu_a, \mu^*)$ to denote the Kullback-Leibler divergence between two Bernoulli distributions with mean $\mu_a$ and $\mu^*$ i.e.,

$$KL(\mu_a, \mu^*) = \mu_a \log \frac{\mu_a}{\mu^*} + (1 - \mu_a) \log \frac{1 - \mu_a}{1 - \mu^*}$$

## 2.2 Structural Causal Model.

The basic semantical framework of our analysis rests on structural causal models (SCM), defined as follows:

DEFINITION 1 (SCM [22]). *A structural causal model (SCM) $M$ is a 4-tuple $\langle U, V, F, P(U)\rangle$ where:*

1. *$U$ is a set of background variables (also called exogenous or latent), that are determined by factors outside of the model;*
2. *$V$ is a set $\{V_1, V_2, \dots, V_n\}$ of observable variables (also called endogenous) that are determined by variables in the model (i.e., by the variables in $U \cup V$);*
3. *$F$ is a set of structural functions $\{f_1, f_2, \dots, f_n\}$ such that each $f_i$ is a mapping from the respective domain of $U_i \cup PA_i$ to $V_i$, where $U_i \subseteq U$ and $PA_i \subseteq V \backslash V_i$ and the entire set $F$ forms a mapping from $U$ to $V$. In other words, each structural function $f_i$, $v_i \leftarrow f_i(pa_i, u_i), i = 1, \dots, n$, assigns a value to $V_i$ that depends on the values of the select set of variables;*
4. *$P(U)$ is a probability distribution over the exogenous variables.*

Each SCM $M$ is associated with a directed acyclic graph (DAG) $G$, where solid nodes correspond to endogenous variables $V$, empty nodes correspond to exogenous variables $U$, and edges represent functional relationships (see Fig. 1 for an example). Let $X$, $Y$, and $Z$ be arbitrary disjoint sets of nodes in a DAG $G$. We denote by $G_{\overline{X}}$ the subgraph obtained by deleting from $G$ all arrows incoming towards $X$, and $G_{\underline{X}}$ the graph obtained by deleting from $G$ all arrows outgoing from $X$. We use $(Y \perp\!\!\!\perp X|Z)_G$ to represent that $X$ is d-separated from $Y$ given $Z$ in graph $G$ – there is no active path between $Y$ and $X$ given $Z$ in graph $G$ [11].

A MAB instance can be naturally encoded as a structural causal model. For example, Fig. 1(a) is the graphical representation of a SCM $M$ for a MAB model where $X$ is the arm selection and $Y$ is the reward variable. We introduce the $do(\cdot)$ operator to differentiate the action $do(X = x)$ from the observation $X = x$ [22, Ch. 3]. In $M$, the expected reward for pulling an arm $x$ is thus the causal effect of an action $do(X = x)$, i.e., $\mu_x = \mathbb{E}[Y|do(X = x)]$. In the structural semantics, actions are modifications of functional relationships. For an arbitrary function $\pi(w)$, action $do(X = \pi(w))$ on a causal model $M$ produces a new model $M_X = \langle U, V, F_X, P(U)\rangle$ where $F_X$ is obtained after replacing $f_i \in F$ for every $V_i \in X$ with the function $\pi$. Fig. 1(b) shows the SCM $M_X$ after action $do(X = x)$ is introduced. With slight abuse of notations, we use $do(x)$ for the actions $do(X = x), x \in D(X)$.

# 3. TRANSFER LEARNING VIA CAUSAL INFERENCE

In this section, we connect causal analysis with transfer learning in RL by discussing two transfer scenarios between MAB agents. We show that when the SCM is provided, the identification algorithm of causal effects can be applied to estimate the effects of an action.

We first consider a transfer scenario between two standard MAB agents $A$ and $A'$ (Fig. 1(b)) that are equipped with the same actuator to perform action $X$. Both agents are unable to observe the context $U$ due to the lack of the corresponding sensor. Let $\epsilon$ be an independent source of randomness.

Agent $A'$ performs an action $do(X = \pi(\epsilon))$ [1] and receives reward $Y$. The experiences collected throughout its interactions with the environment are summarized in the joint distribution $P(x, y)$. Agent $A$, who is observing $A'$ interact with the environment, wants to be more efficient and reuse the observed $P(x, y)$ to find the optimal arm faster. This transfer scenario is summarized as Task I in Fig. 1, where the actions, outcomes, and causal structures used by $A$ and $A'$ coincide. Since the optimal action $do(X = x^*)$ for agent $A$ can be found by evaluating $x^* = \arg\max_{x \in D(X)} \mathbb{E}[Y|do(x)]$, this transfer learning problem can be rephrased as identifying the causal effect $\mathbb{E}[Y|do(x)]$ from the observational distribution $P(x, y)$.

Indeed, this problem has long been studied under the rubric of off-policy learning [32, 25, 24, 17]. The answer in this case is simply to compute the expected conditional reward given $X = x$ based on $P(x, y)$ and use it as if it were the expected reward. Formally, the causal effect can be identified as:

$$\mathbb{E}[Y|do(x)] = \mathbb{E}[Y|x] \tag{1}$$

To prove the above the above statement in causal semantics, we will use the three basic inference rules known as do-calculus [22, Section 3.4.2], which are stated next.

THEOREM 1 (DO-CALCULUS). *Let $G$ be the directed acyclic graph associated with a SCM, and let $P(\cdot)$ stand for the probability distributions induced by that structural model. For any disjoint subsets of variables $X, Y, Z$ and $W$, the following rules hold:*

1. $P(y|do(x), z, w) = P(y|do(x), w)$ *If* $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}$
2. $P(y|do(x), do(z), w) = P(Y|do(X), Z, W)$ *If* $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}$
3. $P(y|do(x), do(z), w) = P(Y|do(X), W)$ *If* $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\overline{Z(W)}}}$

*where $Z(W)$ is the set of $Z$-nodes that are not ancestors [2] of any $W$-node in $G_{\overline{X}}$.*

A target causal effect is identified by applying a sequence of do-calculus rules combined with probabilistic manipulations. The derivation process terminates either when the resulting representation does not contain the $do(\cdot)$ operator or when all possible operations have been exhausted. In latter case, the causal effect is not identifiable (i.e., provably not computable from the available datasets). We will show that Eqn. 1 can be easily obtained through do-calculus. In the DAG $G$ in Fig. 1(b), we note that $X$ is disconnected from $Y$ in the subgraph $G_{\underline{X}}$ (where the outgoing arrows from $X$ are cut), which by rule 2 of do-calculus implies that $P(y|do(x)) = P(y|x)$, therefore $\mathbb{E}[Y|do(x)] = \mathbb{E}[Y|x]$.

Next, we consider a more challenging scenario involving the transfer from a contextual bandit agent $B$ and a standard MAB agent $A$. A contextual bandit is a variation of a MAB where the agent can observe an extra information associated with the reward signal [14]. Agent $B$ (Fig. 1(a)) has the same actuator as $A$ (Fig. 1(b)), but is equipped with an advanced sensor that allows it to observe the context $U$. Agent $B$ observes context $u \in D(U)$, performs action $do(X = \pi(\epsilon, u))$, and receives reward $Y$. The experiences collected throughout its interactions with the environment are summarized by the joint distribution $P(x, y, u)$. The goal of agent $A$ is to find the optimal arm choice based on $P(x, y, u)$. Fig. 1 summarizes this scenario as Task II. Similarly, this problem can be rephrased as identifying the causal effect $\mathbb{E}[Y|do(x)]$ based on $P(x, y, u)$. By applying do-calculus, we have:

[1] This operation in fact represents a stochastic policy.

[2] $Z$ is $W$'s ancestor iff there is a directed path from $Z$ to $W$.

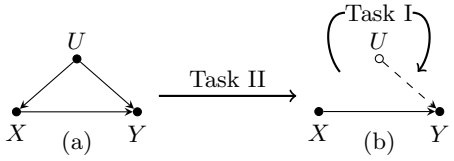



Figure 1: Graphical representation of two identifiable transfer tasks between MABs. (a) the SCM for a contextual bandit agent. (b) the SCM for a standard MAB agent.

$$\begin{aligned}
\mathbb{E}[Y|do(x)] &= \sum_{y \in D(Y)} yP(y|do(x)) \\
&= \sum_{y \in D(Y)} \sum_{u \in D(U)} yP(y|do(x), u)P(u|do(x)) \\
&= \sum_{y \in D(Y)} \sum_{u \in D(U)} yP(y|x, u)P(u|do(x)) \\
&= \sum_{y \in D(Y)} \sum_{u \in D(U)} yP(y|x, u)P(u)
\end{aligned} \tag{2}$$

The last two steps hold by rules 2 and 3, since $Y$ is independent of $X$ given $U$ in the subgraph $G_{\underline{X}}$, and $U$ is independent of $X$ in the subgraph $G_{\overline{U}}$, respectively. Eq. 2 does not contain any $do(\cdot)$ operator. $P(y|x, u)$ and $P(u)$ can both be inferred directly from $P(x, y, u)$. The average effect $\mathbb{E}[Y|do(x)]$ is then said to be identifiable from $P(x, y, u)$.

The two transfer scenarios described above show that the corresponding target effects are identifiable from prior experiences. Indeed, the do-calculus can be applied to any SCM to determine whether a certain causal effects can be identified from the observed data. The procedure can be automated to facilitate the derivation. [31] examines simplified identification criteria and provides an algorithm that was shown to be complete for identification.

# 4. THE CHALLENGES OF NON-IDENTIFIABLE TASKS

While do-calculus provides a systematic way of solving identification queries, it cannot pin down the causal effect for all tasks. When assumptions conveyed by the causal model about the relationship between the sensors and actuators of the source and target MAB agents are not informative enough to supply the missing information of the source distribution, do-calculus fails to reach the transformation formula even after exhausting all possible derivations. In these cases, the causal effect is non-identifiable, i.e., it will not be discernible unambiguously from the given observed behavior of the source agent. In fact, the source distribution and the underlying causal model impose the boundary for the corresponding identification task, which is defined as the problem of identification of causal effects [22, pp. 77]:

DEFINITION 2 (IDENTIFIABILITY). *The causal effect of $X$ on $Y$ is identifiable from $G$ if the quantity $P(y|do(x))$ can be computed uniquely from any positive distribution over the observed variables – that is, if $P_{M_1}(y|do(x)) = P_{M_2}(y|do(x))$*

*for any pair of models $M_1$ and $M_2$ such that $P_{M_1}(v) = P_{M_2}(v) > 0$ and $G(M_1) = G(M_2) = G$.*

To better understand the challenges of non-identifiable tasks, consider again the transfer scenario between a contextual bandit agent $B$ and a standard MAB agent $A$. Instead of receiving the experiences of $B$ in the form of $P(x, y, u)$, assume that $A$ now can only learn by observing $B$'s interactions in the environment. Failing to measure the context $U$ due to lack of a sensor, $A$ can only infer an observational distribution $P(x, y)$. Fig. 2 summarizes this transfer scenario as Task 1. The natural question in this setting is whether the causal effect $\mathbb{E}[Y|do(x)]$ is identifiable from $P(x, y)$. Unfortunately, the answer in this case is negative. To witness, consider a SCM $M$ with $U = (U_1, U_2), X, Y, U_1, U_2 \in \{0, 1\}, X = U_1, Y = X \oplus U_1 \oplus U_2, P(U_1 = 0) = P(U_2 = 0) = 0.1$, where $\oplus$ represents the exclusive-or function. Expected rewards for arm 0 and 1 are respectively $\mu_0 = 0.18, \mu_1 = 0.82$, which implies that the optimal arm is $x^* = 1$. Computing $P(x, y)$ from $M$ leads to $\mathbb{E}[Y|X = 0] = \mathbb{E}[Y|X = 1] = 0.9$. One can now construct another MAB model $M'$ with $U = (U_1, U_2), X, Y, U_1, U_2 \in \{0, 1\}, X = U_1, Y = U_2, P(U_1 = 0) = P(U_2 = 0) = 0.1$. Both $M$ and $M'$ induce the same observational distribution $P(x, y)$, while the expected rewards in $M'$ is $\mu'_0 = \mu'_1 = 0.9$ – the effect is then not identifiable (Def. 2). This means that it is not possible to pin down the average rewards based solely on the available data ($P(x, y)$).

If one is naive about identifiability and transfers $\mathbb{E}[Y|x]$ as if it were $\mathbb{E}[Y|do(x)]$, a negative transfer may occur – i.e., the transferred knowledge will have an adverse impact on the performance of the target agent. Unfortunately, in practice, researchers never have access to the underlying SCM and therefore cannot distinguish the two models. To illustrate this point, we run simulations where 500 samples of $\mathbb{E}[Y|x]$ are naively transferred as if they were $\mathbb{E}[Y|do(x)]$ – see Fig. 3. It is not difficult to note by inspection the significant disparity between the standard Thompson sampling (TS) solver [29, 8] and the Thompson sampling with naive transfer procedure described above ($\text{TS}^-$). In words, $\text{TS}^-$ solver took a significant hit in performance by trying to leverage data collected from the other agent.

In practice, there exist various transfer scenarios where the expected reward cannot be identified. We summarize in Figure 2 and Table 1 three canonical tasks where non-identifiability is an issue. $C_1, C_2$ and $C_3$ represent different conditions (to be defined later on) of prior experiences, which correspond to different (from small to large) bounds over cumulative regret of the target agent. All three tasks are sensible problems which have a wide range of applications in practice. For instance, Task 1, as discussed before, can be seen as the transfer learning problem between a contextual bandit agent and a standard MAB agent with the context unmeasured. Task 2 and Task 3 can be seen as the transfer learning problem between two agents with different actuators, thus having different action spaces [2] (the source agent has a less accurate actuator than the target agent in task 2, and a better one in task 3) . The best possible performance of target agents in Task 1, 2 and 3 are, respectively, regret bounds in case $C_2, C_1$ and $C_3$. We will show in later sections that prior experiences could improve target agent's performance in Task 1 and 2, but not in Task 3.

The lack of identifiability in these three scenarios has been acknowledged in the literature, despite being not necessarily obvious to see by naked eyes. For Task 1, non-identifiability follows due to the unobserved condounding between $X$ and $Y$ discussed in [22, Section 3.5]; non-identifiability from the surrogate variable $Z$ (or, $do(z)$) in Task 2 is more subtle, which was shown in [5]; [23] extended this argument and showed that $\mathbb{E}[Y|do(z)]$ cannot be inferred from $P(y|do(x))$ in Task 3.

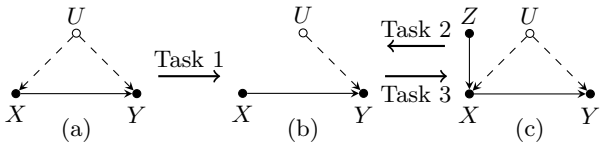



Figure 2: Graphical representation of three canonical transfer tasks where the expected reward is non-identifiable. (a) the SCM for a contextual bandit agent with context $U$ unmeasured. (b) the SCM for a standard MAB agent. (c) the SCM for a standard MAB agent with the action node $Z$.

| Task | Source → Target | | ID | Regret | | |
|---|---|---|---|---|---|---|
| | | | | $C_1$ | $C_2$ | $C_3$ |
| 1 | $P(x, y)$ | $\mathbb{E}[Y\|do(x)]$ | ✗ | ✗ | ✓ | ✓ |
| 2 | $P(x, y\|do(z))$ | $\mathbb{E}[Y\|do(x)]$ | ✗ | ✓ | ✓ | ✓ |
| 3 | $P(z, y\|do(x))$ | $\mathbb{E}[Y\|do(z)]$ | ✗ | ✗ | ✗ | ✓ |

Table 1: Canonical off-policy learning tasks. ID stands for point identifiability. $C_1$, $C_2$, and $C_3$ correspond to the bounds 0, $\mathcal{O}(\log(\log(T)))$ and $\mathcal{O}(\frac{\log(T)}{KL(\mu_a, \mu^*)})$ over the number of draws of sub-optimal arms.

# 5. PRIOR KNOWLEDGE AS BOUNDS

One might surmise that the grim results presented so far imply that no prior data could be useful and experiments should be conducted from scratch, whenever identifiability does not hold. We will show in this section that this is not the case and how prior knowledge, even though imperfect, can help to optimize experimentation. We start by introducing efficient methods to bounding the expected reward of the target quantities listed in Table 1.

We first consider the 2-armed Bernoulli bandits where $X, Y, Z \in \{0, 1\}$. As shown in the previous section, there exists multiple models that agree on the source distribution, but differ wildly on the causal effect. Thus, instead of identifying the exact value of the causal effect, we focus on finding a general set which contains a model compatible with the same source distribution and causal effect. Constraints for the causal effect of every model in the general set must also apply to the target causal effect. [22, Section 8.2] shows a construction of such set for Task 2 by decomposing $U$ into a pair of canonical types $(R_x, R_y)$, where $R_x, R_y \in \{0, 1, 2, 3\}$ and represent the different types of individuals in the population. We extend this idea to bound $\mathbb{E}[Y|do(x)]$ in Task 1. We decompose the latent variable $U$ into a pair of discrete variables $(R_x, R_y)$, where $R_x \in \{0, 1\}, R_y \in \{0, 1, 2, 3\}$. Let $q_{ij} = P(R_x = i, R_y = j) \geq 0$, and $Q = \{q_{ij}\}$. For $\forall x \in D(X), \forall r_x \in D(R_x), \forall r_y \in D(R_y)$, $X$ and $Y$ are decided by functions $X = f_X(r_x)$ and $Y = f_Y(x, r_y)$ defined

as follows:

$$f_X(r_x) = r_x \qquad f_Y(x, r_y) = \begin{cases} 0 & \text{if } r_y = 0 \\ x & \text{if } r_y = 1 \\ 1-x & \text{if } r_y = 2 \\ 1 & \text{if } r_y = 3 \end{cases} \tag{3}$$

The values of $R_y$ represents the canonical types and can be given causal interpretation, namely, "doomed," "helped," "hurt," and "immune" [9] . Let $p_{ij} = P(X = i, Y = j)$. $P(x, y)$ and $\mathbb{E}[Y|do(x)]$ can be written as linear combinations in the space spanned by $Q$:

$$\begin{aligned} p_{00} &= q_{00} + q_{01} & p_{01} &= q_{02} + q_{03} \\ p_{10} &= q_{10} + q_{12} & p_{11} &= q_{11} + q_{13} \end{aligned} \tag{4}$$

$$\mathbb{E}[Y|do(X = 0)] = q_{02} + q_{03} + q_{12} + q_{13} \tag{5}$$

$$\mathbb{E}[Y|do(X = 1)] = q_{01} + q_{03} + q_{11} + q_{13} \tag{6}$$

We can then lower (upper) bound $\mathbb{E}[Y|do(x)]$ by minimizing (maximizing) Eqs. 5 and 6 subject to constraints 4 and $q_{ij} \geq 0$, which leads to a closed-form solution shown next.

THEOREM 2. *Consider Task 1 with $X, Y \in \{0, 1\}$, given $P(x, y)$, $\mathbb{E}[Y|do(x)]$ can be bounded by:*

$$\mathbb{E}[Y|do(X = 0)] \in [p_{01}, p_{01} + p_{10} + p_{11}]$$
$$\mathbb{E}[Y|do(X = 1)] \in [p_{11}, p_{11} + p_{00} + p_{01}]$$

PROOF. Based on constraints 4, we have:

$$\begin{aligned} q_{00} &= p_{00} - q_{01} & q_{02} &= p_{01} - q_{03} \\ q_{10} &= p_{10} - q_{12} & q_{11} &= p_{11} - q_{13} \end{aligned} \tag{7}$$

Since $q_{i,j} \geq 0$, $q_{01}, q_{03}, q_{12}, q_{13}$ are independent variables taking values in:

$$\begin{aligned} q_{01} &\in [0, p_{00}] & q_{03} &\in [0, p_{01}] \\ q_{12} &\in [0, p_{10}] & q_{13} &\in [0, p_{11}] \end{aligned}$$

Replace $q_{00}, q_{02}, q_{10}, q_{11}$ in Equation 5 and 6 with Equations 7, we have:

$$\mathbb{E}[Y|do(X = 0)] = p_{01} + q_{12} + q_{13} \in [p_{01}, p_{01} + p_{10} + p_{11}]$$
$$\mathbb{E}[Y|do(X = 1)] = p_{11} + q_{01} + q_{03} \in [p_{11}, p_{11} + p_{00} + p_{01}]$$

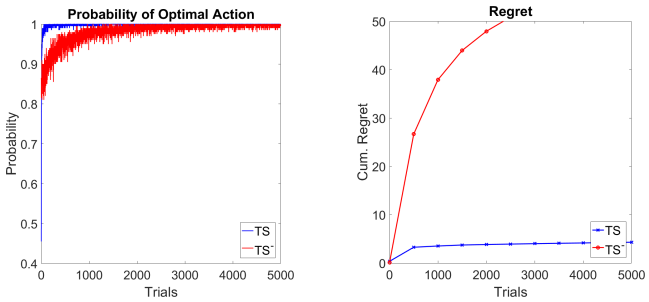
□

From a causal perspective, this simple bound is unexpected since it was not believed that a model without any independencies or exclusion restrictions could impose any informative constraint over the experimental distribution.

Considering the task 3, we decompose and discretize $U$ into a tuple $(R_z, R_x, R_y)$, where $R_z \in \{0, 1\}, R_x, R_y \in \{0, 1, 2, 3\}$. Let $q_{ijk} = P(R_z = i, R_x = j, R_y = k) \geq 0$. $Y$ is decided by the function $f_Y(x, r_y)$ defined in Eq. 3. For $\forall z \in D(Z), \forall r_x \in D(R_x), \forall r_z \in D(R_z)$, $X, Z$ are decided by functions $X = f_X(z, r_x)$ and $Z = f_Z(r_z)$ defined as follows:

$$f_X(z, r_x) = \begin{cases} 0 & \text{if } r_x = 0 \\ z & \text{if } r_x = 1 \\ 1-z & \text{if } r_x = 2 \\ 1 & \text{if } r_x = 3 \end{cases} \qquad f_Z(r_z) = r_z$$

Let $p_{ijk} = P(Z = i, Y = j|do(X = k))$. Again, we can write $P(z, y|do(x))$ and $\mathbb{E}[Y|do(z)]$ as linear combinations of




(a) (b)

Figure 3: Simulation results of negative transfer example in Sec. 4 comparing the standard Thompson sampling (TS) and Thompson sampling with naive transfer procedure ($TS^-$).

$q_{ijk}$. The lower (upper) bound can then be obtained by minimizing (maximizing) $\mathbb{E}[Y|do(z)]$ subject to the constraints imposed by $P(z, y|do(x))$ and $q_{ijk} \geq 0$, which leads to a closed-form solution shown in the theorem below.

THEOREM 3. *Consider Task 3 with $X, Y, Z \in \{0, 1\}$, given $P(z, y|do(x))$, $\mathbb{E}[Y|do(z)]$ can be bounded by:*

$$\mathbb{E}[Y|do(Z = 0)] \in [l, h], \quad \mathbb{E}[Y|do(Z = 1)] \in [l, h], \quad \textit{where:}$$

$$l = \max \left\{ \begin{array}{l} 0 \\ p_{001} + p_{110} + p_{011} - p_{000} - p_{010} - p_{000} \\ p_{010} - p_{001} \\ p_{011} + p_{110} - p_{000} - p_{101} \end{array} \right\}$$

$$h = \min \left\{ \begin{array}{l} p_{001} + p_{100} + 2p_{011} + 2p_{110} - p_{000} - p_{101} \\ p_{010} + p_{100} + p_{110} + p_{011} \\ p_{100} + 2p_{110} + 2p_{001} + 2p_{011} - p_{000} - p_{010} - p_{101} \\ p_{001} + p_{011} + p_{100} + p_{010} \end{array} \right\}$$

Even though embedded in a more constrained structure, the bounds of Thm. 3 are weaker than Thm. 2 since both arms coincide. [23] showed that identification of $do(Z)$ is infeasible from experiments over $X$ in task 3, and Thm. 3 provides a stronger condition saying that not even an informative bound can be derived in such settings.

The construction given above can be extended to any discrete setting where $X, Y, Z$ have larger dimensions. For Task 1 and 2, $U$ should be decomposed and discretized into a pair $(R_x, R_y)$, where $|R_x| = |X|, |R_y| = |Y|^{|X|}$ in task 1, and $|R_x| = |X|^{|Z|}, |R_y| = |Y|^{|X|}$ in task 2. For task 3, $U$ is decomposed into a tuple $(R_z, R_x, R_y)$, where $|R_z| = |Z|, |R_x| = |X|^{|Z|}, |R_y| = |Y|^{|X|}$. The necessary distributions can be written as linear combinations in the corresponding $Q$-space. The bounds can then be derived by solving a series of linear optimization problems using a standard LP solver.

## 6. MULTI-ARMED BANDITS WITH CAUSAL BOUNDS

We discuss in this section how the causal bounds can be used to more efficiently identify an optimal treatment arm. We consider an augmented stochastic MAB problem with a prior represented as a list of bounds over expected rewards. Formally, for any arm $a$, let $[l_a, h_a]$ be the bound for $\mu_a$ such that $\mu_a \in [l_a, h_a]$. Without loss of generality, we assume $0 < l_a < h_a < 1$ and denote by $l_{max}$ the maximum of all

**Algorithm 1:** The kl-UCB algorithm with bounds over expected reward (B-kl-UCB)

**Input:** A non-decreasing function $f : \mathbb{N} \to \mathbb{R}$
A list of bounds for expected rewards $\{[l_1, h_1], \ldots, [l_K, h_K]\}$
**Initialization:** Remove any arm $a$ with $h_a < l_{max}$.
Let $K'$ denote the number of remaining arms.
Pull each arm of $\{1, \ldots, K'\}$ once
**for all** $t = K'$ to $T - 1$ **do**
  Compute for each arm $a$ the quantity:

$$\hat{U}_a(t) = \min\left\{U_a(t), h_a\right\}$$

  where

$$U_a(t) = \sup\left\{\mu \in [0,1] : KL(\hat{\mu}_a(t), \mu) \leq \frac{f(t)}{N_a(t)}\right\}$$

  Pick an arm $A_t = \arg\max_{a \in \{1,\ldots,K'\}} \hat{U}_a(t)$.
**end for**

**Algorithm 2:** Thompson Sampling for Bernoulli bandit with bounds over expected reward (B-TS)

**Input:** A non-decreasing function $f : \mathbb{N} \to \mathbb{R}$
A list of bounds for expected rewards $\{[l_1, h_1], \ldots, [l_K, h_K]\}$
**Initialization:** Remove any arm $a$ with $h_a < l_{max}$.
Let $K'$ denote the number of remaining arms.
$S_a = 0, F_a = 0, \forall a \in \{1, \ldots, K'\}$
**for all** $t = 0$ to $T - 1$ **do**
  **for all** $a = 1$ to $K'$ **do**
    **repeat**
      Draw $\theta_a \sim Beta(S_a + 1, F_a + 1)$
    **until** $\theta_a \in [l_a, h_a]$
  **end for**
  Draw arm $A_t = \arg\max_{a \in \{1,\ldots,K'\}} \theta_a$ and observe reward $y$.
  **if** $y = 1$ **then** $S_i = S_i + 1$
  **else** $F_i = F_i + 1$
**end for**

lower bounds, i.e., $l_{max} = \max_{i=1,\ldots,K} l_i$. Let $h^* = h_{a^*}$ and $l^* = l_{a^*}$.

UCB constitutes an elegant family of algorithms that has been used in a number of settings given its attractive guarantees – its regret grows only logarithmically with the number of actions taken [3, 7]. We augment UCB to take into account the causal bounds, which we call B-kl-UCB (Algorithm 1). B-kl-UCB exploits the causal bound in two ways: 1. filter any arm $a$ during initialization if $h_a < l_{max}$; 2. truncate the UCB $U_a(t)$ with $\hat{U}_a(t) = \min\{U_a(t), h_a\}$ and picks an arm with the largest $\hat{U}_a(t)$. To understand the implications of these modifications, we derive in the sequel the corresponding regret bound.

THEOREM 4. *Consider a $K$-MAB problem with rewards bounded in $[0, 1]$. For each arm $a \in \{1, \ldots, K\}$ and expected reward $\mu_a$ bounded by $[l_a, h_a]$, where $0 < l_a < h_a < 1$. Choosing the parameters $f(t) = \log(t) + 3\log(\log(t))$, in B-kl-UCB algorithm, the number of draws any sub-optimal arm $a$ is upper bounded for any horizon $T \geq 3$ as:*

$$\mathbb{E}[N_a(T)] \leq \begin{cases} 0 & \text{if } h_a < l_{max} \\ 4 + 4e\log(\log(T)) & \text{if } l_{max} \leq h_a < \mu^* \\ \frac{\log(T)}{KL(\mu_a, \mu^*)} + \mathcal{O}(\frac{\log(\log(T))}{KL(\mu_a, \mu^*)}) & \text{if } h_a \geq \mu^* \end{cases}$$

To prove Thm.4, we first introduce two lemmas:

LEMMA 1. *Consider a $K$-armed bandit problem and $f(t)$ defined in Theorem 4. In B-kl-UCB algorithm, the term $\sum_{t=K'}^{T-1} \mathbb{P}\{\hat{U}_{a^*}(t) < \mu^*\}$ is bounded by:*

$$\sum_{t=K'}^{T-1} \mathbb{P}(\hat{U}_{a^*}(t) < \mu^*) \leq 3 + 4e\log(\log(T))$$

PROOF. Since $\hat{U}_{a^*}(t) = \min\left\{U_{a^*}(t), h^*\right\}$, the means $\mu^*$ is larger than either $U_{a^*}(t)$ or $h^*$. Thus, we have:

$$\begin{aligned} \sum_{t=K'}^{T-1} \mathbb{P}(\hat{U}_{a^*}(t) < \mu^*) &\leq \sum_{t=K'}^{T-1} \mathbb{P}(U_{a^*}(t) < \mu^*) + \sum_{t=K'}^{T-1} \mathbb{P}(h^* < \mu^*) \\ &= \sum_{t=K'}^{T-1} \mathbb{P}(U_{a^*}(t) < \mu^*) \quad \text{By definition } h^* \geq \mu^* \end{aligned}$$

By [7, Fact A.1], we have:

$$\sum_{t=K'}^{T-1} \mathbb{P}(\hat{U}_{a^*}(t) < \mu^*) \leq \sum_{t=K'}^{T-1} \mathbb{P}(U_{a^*}(t) < \mu^*) \leq 3 + 4e\log(\log(T))$$

□

LEMMA 2. *Consider the $K$-armed bandit problem and $f(t)$ defined in Theorem 4. In B-kl-UCB algorithm, the term $\sum_{t=K'}^{T-1} \mathbb{P}(\mu^* \leq \hat{U}_a(t), X_t = a)$ is bounded by:*

$$\begin{cases} 0 & \text{if } l_{max} \leq h_a < \mu^* \\ \frac{\log(T)}{KL(\mu_a, \mu^*)} + \mathcal{O}(\frac{\log(\log(T))}{KL(\mu_a, \mu^*)}) & \text{if } h_a \geq \mu^* \end{cases}$$

PROOF. For all $n > 1$, let $\hat{\mu}_a(t)$ be the empirical estimation of $\mu_a$, and $\tau_{a,n}$ denote the round at which $a$ was pulled for the $n$-th time, For reward samples from $\nu_a$, $\{Y_{a,0}, \ldots, Y_{a,n}\}$, define $\hat{\mu}_{a,n} = \frac{1}{n}\sum_{s=1}^{n} Y_{a,s}$. We of course have the writing $\hat{\mu}_a(t) = \hat{\mu}_{a,N_a(t)}$. We now bound the term by cases:

- **Case 1.** $h_a < \mu^*$. Since $\hat{U}_a(t) \leq h_a$, we must have $\mu^* \leq U_a(t) \leq h_a$ which contradicts the fact $h_a < \mu^*$. This means that $\sum_{t=K'}^{T-1} \mathbb{P}(\mu^* \leq \hat{U}_a(t), X_t = a) = 0$.
- **Case 2.** $h_a \geq \mu^*$. Since $\hat{U}_{a^*}(t) = \min\left\{U_{a^*}(t), h^*\right\}$, this means $\mu^*$ must be upper bounded by both $U_{a^*}(t)$ and $h^*$. Thus, we have:

$$\begin{aligned} &\sum_{t=K'}^{T-1} \mathbb{P}(\mu^* \leq \hat{U}_a(t)) \leq \sum_{t=K'}^{T-1} \mathbb{P}(\mu^* \leq U_a(t), \mu^* \leq h_a, X_t = a) \\ &= \sum_{t=K'}^{T-1} \mathbb{P}(\mu^* \leq U_a(t), X_t = a) \quad \text{By definition } h_a \geq \mu^* \\ &= \sum_{t=K'}^{T-1} \mathbb{P}(\exists \mu \in [\mu^*, 1] : KL(\hat{\mu}_a(t), \mu) \leq \frac{f(t)}{N_a(t)}, X_t = a) \\ &= \sum_{n=1}^{T-K'} \sum_{t=\tau_{a,n}+1}^{\tau_{a,n+1}} \mathbb{P}(\exists \mu \in [\mu^*, 1] : KL(\hat{\mu}_{a,n}, \mu) \leq \frac{f(t)}{n}, X_t = a) \\ &\leq \sum_{n=1}^{T-K'} \sum_{t=\tau_{a,n}+1}^{\tau_{a,n+1}} \mathbb{P}(\exists \mu \in [\mu^*, 1] : KL(\hat{\mu}_{a,n}, \mu) \leq \frac{f(T)}{n}, X_t = a) \end{aligned}$$

$$= \sum_{n=1}^{T-K'} \mathbb{P}(\exists \mu \in [\mu^*, 1] : KL(\hat{\mu}_{a,n}, \mu) \leq \frac{f(T)}{n})$$
$$\leq n_0 + \sum_{n=n_0+1}^{T-K'} \mathbb{P}(\exists \mu \in [\mu^*, 1] : KL(\hat{\mu}_{a,n}, \mu) \leq \frac{f(T)}{n})$$

where $n_0 = \lceil \frac{f(T)}{KL(\mu_a, \mu^*)} \rceil$. This implies

$$(\forall n \geq n_0 + 1) \quad KL(\mu_a, \mu^*) > \frac{f(T)}{n}$$

Since $KL(\cdot, \mu^*)$ is continuous decreasing function on $[0, \mu^*]$, there must $\exists \mu_{\frac{f(T)}{n}} \in (\mu_a, \mu^*]$, such that:

$$KL(\mu_{\frac{f(T)}{n}}, \mu^*) \geq \frac{f(T)}{n}$$

We next show that:

$$\{\exists \mu \in [\mu^*, 1] : KL(\hat{\mu}_{a,n}, \mu) \leq \frac{f(T)}{n}\} \Rightarrow \{\hat{\mu}_{a,n} \geq \mu_{\frac{f(T)}{n}}\}$$

This can be proved by contradiction. Suppose $\hat{\mu}_{a,n} < \mu_{\frac{f(T)}{n}}$, we then have:

$$(\forall \mu \in [\mu^*, 1]) \quad KL(\hat{\mu}_{a,n}, \mu) \geq KL(\hat{\mu}_{a,n}, \mu^*)$$
$$> KL(\mu_{\frac{f(T)}{n}}, \mu^*) = \frac{f(T)}{n}$$

which contradicts $\{\exists \mu \in [\mu^*, 1] : KL(\hat{\mu}_{a,n}, \mu) \leq \frac{f(T)}{n}\}$. Thus, $\forall \lambda > 0$, we have:

$$\mathbb{P}(\exists \mu \in [\mu^*, 1] : KL(\hat{\mu}_{a,n}, \mu) \leq \frac{f(T)}{n})$$
$$\leq \mathbb{P}(\hat{\mu}_{a,n} \geq \mu_{\frac{f(T)}{n}}) \leq e^{-\lambda \mu_{\frac{f(T)}{n}}} \mathbb{E}[e^{\lambda \hat{\mu}_{a,n}}]$$

By [7, Fact A.2], we have:

$$\sum_{t=K'}^{T-1} \mathbb{P}(\mu^* \leq \hat{U}_a(t)) \leq \frac{\log(T)}{KL(\mu_a, \mu^*)} + \mathcal{O}(\frac{\log(\log(T))}{KL(\mu_a, \mu^*)})$$

□

We now proceed to complete the proof of Thm.4.

PROOF (**Proof of Theorem 4**). Without loss of generality, let $K' \geq 2$. The proof for the case $h_a < l_{max}$ is trivial, since arms satisfying this condition are removed in the initialization and never played. We next focus on the other two cases. By definition of the algorithm, at rounds $t \geq K'$, one has $X_{t+1} = a$ only if $U_a(t) \geq U_{a^*}(t)$. Therefore, we can follow the same decomposition in [7]:

$$\{X_t = a\} \subseteq \{\hat{U}_{a^*}(t) < \mu^*\} \cup \{\mu^* \leq \hat{U}_{a^*}(t), X_t = a\}$$
$$\subseteq \{\hat{U}_{a^*}(t) < \mu^*\} \cup \{\mu^* \leq \hat{U}_a(t), X_t = a\} \quad (8)$$

Then, the expected number of trial for arm $a$ after $T$ rounds, $\mathbb{E}[N_a(T)]$, can be rewritten as:

$$\mathbb{E}[N_a(T)] = 1 + \underbrace{\sum_{t=K'}^{T-1} \mathbb{P}(\hat{U}_{a^*}(t) < \mu^*)}_{\text{Term (1)}} + \underbrace{\sum_{t=K'}^{T-1} \mathbb{P}(\mu^* \leq \hat{U}_a(t)), X_t = a)}_{\text{Term (2)}}$$

Term 1 and 2 are bounded by Lemma 1 and 2 respectively. Put everything together, we prove the statement. □

Thm.4 demonstrates the potential improvements due to the causal bounds. If the causal bounds impose strong constraints over the arm's distribution, B-kl-UCB provides asymptotic improvements over kl-UCB, and essentially dominates kl-UCB. On the other hand, when such constraints are too weak ($h_a \geq \mu^*$), the bound of B-kl-UCB degenerates to the standard kl-UCB bound. This latter bound is not improvable in a sense that there exists a parametrization such that for any admissible strategy (not grossly under-performing), the regret is bounded below by $\sum_{a:\mu_a<\mu^*} \frac{\log(T)}{KL(\mu_a,\mu^*)}$.

Let $h_a < l_{max}$, $l_{max} \leq h_a < \mu^*$ and $h_a \geq \mu^*$ be denoted by cases $C_1$, $C_2$, and $C_3$. Thm. 4 indicates that for $C_1$ and $C_2$, the number of draws in B-kl-UCB for any sub-optimal arm is bounded by (respectively) 0 and $\mathcal{O}(\log(\log(T))$, which is not achievable by kl-UCB. For $C_3$, B-kl-UCB and kl-UCB obtains the same bound with $\mathcal{O}(\frac{\log(T)}{KL(\mu_a,\mu^*)})$. We next prove that the bound for $C_3$ cannot be improved.

THEOREM 5. *(**Lower Bound for** $h_a \geq \mu^*$) Consider a strategy that satisfies $\mathbb{E}[T_i(n)] = o(n^\alpha)$ for any Bernoulli distribution, any arm $i$ with $\Delta_i > 0$, and any $\alpha > 0$. Then, for any arm $i$ with $\Delta_i > 0$ and $h_a > \mu^*$, the following holds*

$$\liminf_{n \to +\infty} \frac{E[N_i(n)]}{\log(n)} \geq \frac{1}{KL(\mu_i, \mu^*)}$$

PROOF. Without loss of generality, let $i = 1$ with $\mu_1 > \mu^*$ and $\mu_2 = \mu^*$. Since $KL(\mu_1, \cdot)$ is a continuous function, for any $\epsilon > 0$, there exists $\mu_1' \in (\mu^*, h_1]$ such that:

$$KL(\mu_1, \mu_1') \leq (1 + \epsilon) KL(\mu_1, \mu^*)$$

We now have two bandit parameter vectors $(\mu_1, \mu^*, \ldots, \mu_k)$ and $(\mu_1', \mu^*, \ldots, \mu_k)$. Let $0 < \alpha < \epsilon$, and $C_n$ denote the event:

$$C_n = \left\{ N_1(n) < \frac{(1-\epsilon)}{KL(\mu_1, \mu_1')}, \hat{kl}_{N_1(n)} \leq (1-\alpha)\log(n) \right\}$$

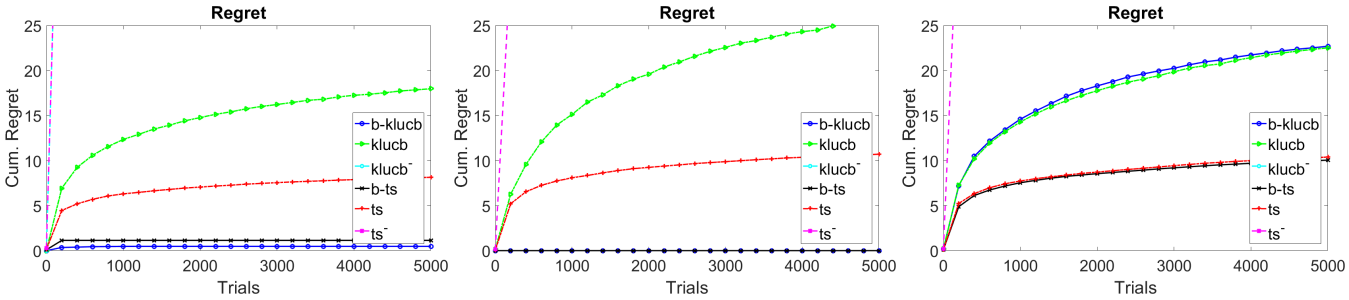
where $\hat{kl}_m$ is defined as

$$\hat{kl}_m = \sum_{t=1}^{m} \log \frac{\mu_1 Y_{1,t} + (1-\mu_1)(1-Y_{1,t})}{\mu_1' Y_{1,t} + (1-\mu_1')(1-Y_{1,t})}$$

The rest follows the proof of [13, Theorem 2]. Q.E.D. □

**Remark (revisiting the three canonical tasks)** These results imply that the constraints imposed by the bounds over the expected rewards $(C_1, C_2, C_3)$ translate into different regret bounds for the MAB agent. Interestingly, a simple analysis reveals that the canonical tasks can be associated with these different cases, which is summarized in Table 1. We can see that there exist no parametrization for task 1 satisfying $C_1$ since $p_{0,1} \leq p_{1,1}+p_{0,0}+p_{0,1}, p_{1,1} \leq p_{0,1}+p_{1,0}+p_{1,1}$ (by Thm. 2). Still, there exists some instances falling into $C_2, C_3$. Considering the bounds implied by task 2 [22, pp. 250], we can see that there exist a number of instances compatible with $C_1$ and $C_2$ (e.g., pick large values for $p_{000}$ and $p_{111}$), which means that there is great potential for improvement in this task. For task 3, Thm. 3 indicates that the bounds must be the same for both arms, which rules out $C_1$ and $C_2$, and makes the task to fall into $C_3$. Since this case is not improvable by Thm. 5, this is less interesting among the canonical problems.

There exists a class of algorithms based on Thompson sampling (TS) [29, 8] that presents strong empirical performance, but its theoretical analysis was not completely




(a) Task 1 with parametrizations in $C_2$ (b) Task 2 with parametrizations in $C_1$ (c) Task 3 with parametrizations in $C_3$

Figure 4: Simulations results of the canonical tasks (Table 1) comparing solvers that are causal enhanced (B-kl-UCB, B-TS), standard (kl-UCB, TS), and naive (kl-UCB$^-$, TS$^-$). Graphs are rendered in high resolution and can be zoomed in.

understood until recently [1]. Given that kl-UCB and B-kl-UCB are asymptotically equivalent for parametrizations in $C_3$, we augment a basic TS solver to consider the causal bounds, which we call B-TS (Algorithm 2). We employ a "rejection sampling" approach to enforce the causal bound $[l_a, h_a]$ on the estimated expected reward $\Theta_a$. Simulation results will be discussed next and compare the performance of B-TS and B-kl-UCB. We note that the regret analysis of B-TS is a challenging and still open problem.

## 7. EXPERIMENTAL RESULTS

In this section, we conduct experiments to validate our findings. In particular, we compare B-kl-UCB and B-TS against standard MAB algorithms (kl-UCB and TS) that have no access to the causal bounds. We also include their counterparts that incorporate the data from the source agent using the naive transfer procedure described in Sec. 4 (without distinguishing the observational and experimental distributions), which we call kl-UCB$^-$ and TS$^-$. We presents simulation results for 2-armed Bernoulli bandits.

The simulations are partitioned into rounds of $T = 5000$ trials averaged over $N = 200$ repetitions. For each task, we collect 5000 samples generated by a source agent and compute the empirical source distribution. The causal bounds are estimated from the empirical distribution with the methods described in Sec. 5. We assess each algorithm's performance in terms of their cumulative regrets (CR).

**Task 1.** The expected rewards of the given parametrization are $\mu_1 = 0.66, \mu_2 = 0.36$, and the estimated causal bounds are $b_1 = [0.03, 0.76], b_2 = [0.21, 0.51]$. Since $h_2 < \mu^* = \mu_1$, this parametrization satisfies condition $C_2$. The results are shown in Fig. 4a and reveal a significant difference in the regret experienced by B-kl-UCB (CR = 0.47) and B-TS (CR = 1.14) compared to kl-UCB (CR = 17.97) and TS (CR = 8.14). kl-UCB$^-$ (CR = 1499.70) and TS$^-$ (1499.99) perform worst among all strategies due to the negative transfer.

**Task 2.** The expected rewards of the given parametrization are $\mu_1 = 0.58, \mu_2 = 0.74$ and the estimated causal bounds are $b_1 = [0.48, 0.61], b_2 = [0.7, 0.83]$. Since $h_1 < l_{max} = l_2$, this parametrization falls into $C_1$'s bucket. Fig. 4b reveals a significant difference in the regret experienced by B-kl-UCB (CR = 0.00) and B-TS (CR = 0.00) compared to kl-UCB (CR = 25.94) and TS (CR = 10.70). kl-UCB$^-$ (CR = 799.84) and TS$^-$ (CR = 800.00) perform worst among all strategies due to the negative transfer of samples.

**Task 3.** The expected rewards are $\mu_1 = 0.2, \mu_2 = 0.4$ and the estimated causal bounds are $b_1 = b_2 = [0, 0.61]$. Since $h_1 \geq \mu^* = \mu_2$, this parametrization satisfies $C_3$. Simulation results shown in Fig. 4c reveal minor differences in the regret experienced by B-kl-UCB (CR = 23.70) and B-TS (CR = 10.05) compared to their counterparts kl-UCB (CR = 22.51) and TS (CR = 10.40). Again, kl-UCB$^-$ (CR = 999.8) and TS$^-$ (CR = 1000.00) perform worst among all strategies due to the negative transfer of samples.

These results corroborate with the theoretical results in the previous section and show that prior experiences can be transferred to improve the performance of the target agent, even when identifiability does not hold. Specifically, B-kl-UCB dominates kl-UCB in $C_1, C_2$ while obtaining a similar performance in $C_3$. Interestingly, B-TS exhibits similar behavior as B-kl-UCB in $C_1, C_2$, and superior performance (with TS) in $C_3$. This suggests that B-TS is an attractive practical alternative when causal bounds are available, despite its lack of theoretical guarantees. Clearly, a naive treatment of previously collected samples causes a negative transfer and deterioration of the algorithm's performance.

## 8. CONCLUSION

We considered in this paper the transfer learning problem between two MAB agents when the causal model of the environment is provided. We show that the identification algorithm of causal inference can be applied to estimate the expected reward of the target agent. When the expected reward is not identifiable, partial information can still be extracted as causal bounds to improve the performance of the solver. Our technique provides the basis for principled transfer learning when the standard assumptions underlying off-policy learning are violated.

There are interesting research questions opened by this paper. First, simulation results suggest that, in practice, B-TS (the sampling-based algorithm with causal bounds) is preferred over UCB-type algorithms. The theoretical analysis for B-TS is still not fully understood. We conjecture that the regret bound of B-TS will be asymptotically the same as B-kl-UCB. Second, causal models can be applied to more general RL settings (e.g., MDPs, POMDPs) such that causal effects across distinct domains can be leveraged. When target effects are not identifiable, it is possible to learn partial information about the target task from source data, and partial information could be used to speed-up learning.

## REFERENCES


[1] S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. *CoRR*, abs/1111.1797, 2011.

[2] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[4] E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2015.

[5] E. Bareinboim and J. Pearl. Causal inference by surrogate experiments: $z$-identifiability. In N. de Freitas and K. Murphy, editors, *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 113–120, Corvallis, OR, 2012. AUAI Press.

[6] E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113:7345–7352, 2016.

[7] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, G. Stoltz, et al. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

[8] O. Chapelle and L. Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[9] D. Heckerman and R. Shachter. A definition and graphical representation for causality. In P. Besnard and S. Hanks, editors, *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 262–273, San Francisco, 1995. Morgan Kaufmann.

[10] Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. In R. Dechter and T. Richardson, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 217–224. AUAI Press, Corvallis, OR, 2006.

[11] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[12] G. Konidaris and A. G. Barto. Building portable options: Skill transfer in reinforcement learning.

[13] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[14] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

[15] A. Lazaric. Transfer in reinforcement learning: a framework and a survey. In *Reinforcement Learning*, pages 143–173. Springer, 2012.

[16] Y. Liu and P. Stone. Value-function-based transfer for reinforcement learning using structure mapping. 2006.

[17] H. R. Maei, C. Szepesvári, S. Bhatnagar, and R. S. Sutton. Toward off-policy learning control with function approximation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 719–726, 2010.

[18] N. Mehta, S. Natarajan, P. Tadepalli, and A. Fern. Transfer in variable-reward hierarchical reinforcement learning. *Machine Learning*, 73(3):289–312, 2008.

[19] N. Mehta, S. Ray, P. Tadepalli, and T. Dietterich. Automatic discovery and transfer of maxq hierarchies. In *Proceedings of the 25th international conference on Machine learning*, pages 648–655. ACM, 2008.

[20] N. Mehta, S. Ray, P. Tadepalli, and T. Dietterich. Automatic discovery and transfer of task hierarchies in reinforcement learning. 2011.

[21] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 1995.

[22] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. 2nd edition, 2009.

[23] J. Pearl. Is scientific knowledge useful for policy analysis? a peculiar theorem says: No. *Journal of Causal Inference J. Causal Infer.*, 2(1):109–112, 2014.

[24] D. Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[25] D. Precup, R. S. Sutton, and S. Dasgupta. Off-policy temporal-difference learning with function approximation. In *ICML*, pages 417–424, 2001.

[26] H. Robbins. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pages 169–177. Springer, 1985.

[27] I. Shpitser and J. Pearl. Identification of conditional interventional distributions. In R. Dechter and T. Richardson, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 437–444. AUAI Press, Corvallis, OR, 2006.

[28] M. E. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009.

[29] W. 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.

[30] J. Tian and J. Pearl. A general identification condition for causal effects. Technical Report R-290-A, Department of Computer Science, University of California, Los Angeles, CA, 2003.

[31] J. Tian and J. Pearl. On the identification of causal effects. Technical Report R-290-L, Department of Computer Science, University of California, Los Angeles, CA, 2003.

[32] C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.