Transfer Learning in Multi-Armed Bandits:
A Causal Approach

Junzhe Zhang and **Elias Bareinboim**

Purdue University, USA

{zhang745,eb}@purdue.edu

Abstract

Reinforcement learning (RL) agents have been deployed in complex environments where interactions are costly, and learning is usually slow. One prominent task in these settings is to reuse interactions performed by other agents to accelerate the learning process. Causal inference provides a family of methods to infer 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 do-calculus [Pearl, 2000] and standard learning techniques. Our new identification strategy combines two steps – first, deriving bounds over the arms 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 achieves orders of magnitude faster convergence rates than these algorithms. Finally, we perform simulations and empirically demonstrate that our strategy is consistently more efficient than the current (non-causal) state-of-the-art methods.

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 [Konidaris and Barto, 2007; Liu and Stone, 2006] (for a survey, see [Lazaric, 2012; Taylor and Stone, 2009]).

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 (source) [Pearl, 2000; Bareinboim and Pearl, 2016]. 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). Qualitative knowledge about causal 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 of users click-through rate causes an observed revenue growth of an advertising engine, not the other way around.

In his seminal work, [Pearl, 1995] 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 identification, i.e., any causal effect can be identified from observational and experimental data if and only if it can be derived by do-calculus [Tian and Pearl, 2002; Shpitser and Pearl, 2006; Huang and Valtorta, 2006; Bareinboim and Pearl, 2012].

Despite its success in identifying the effect of actions from heterogeneous data in compelling settings across the sciences [Bareinboim and Pearl, 2016], causal inference techniques have rarely been used to assist the transfer of knowledge in interactive domains. [Mehta *et al.*, 2008; 2011] 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 did not considered the transfer of causal knowledge. Connections between causal models with UCs and RL were first established in [Bareinboim *et al.*, 2015]. 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 identify the causal effect of the target agent's action from trajectories of the source agent. In particular, we study three canonical settings where the effect is non-identifiable 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 and off-policy evaluation.

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 [Cappé *et al.*, 2013]. 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 algorithms (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 and Notations

In this section, we introduce the basic notations and definitions used throughout the paper. We use the capital letter X for the variable name and lowercase letter x for a specific value taken by X. Let $D(X)$ and $|X|$ denote by, repsectively, the domain and the dimension of variable X. We will consistently use the abbreviation $P(x)$ for the probabilities $P(X = x), x \in D(X)$.

2.1   Structural Causal Model

We will use structural causal models (SCMs) [Pearl, 2000, Ch. 7] as the basic semantical framework of our analysis. A SCM M consists of a set of observed (endogenous) variables V and unobserved (exogenous) variables U. The values of each endogenous variable $V_i \in V$ are determined by a structural function f_i taking as argument a combination of the other endogenous and exogenous variables (i.e., $V_i \leftarrow f_i(PA_i, U_i), Pa_i \subseteq V, U_i \subseteq U)$. The values of the exogenous variables U are drawn from a distribution $P(U)$. A causal diagram associated with the SCM M is a directed acyclic graph where solid nodes correspond to endogenous variables (V), empty nodes correspond to exogenous variables (U), and edges represent causal relations (see Fig. 1).

An attractive feature of SCMs is that they are capable of representing causal operations such as interventions, in addition to standard probabilistic operations such as marginalization and conditioning. The $do(\cdot)$ operator is used to denote interventions (actions) [Pearl, 2000, Ch. 3]. For an arbitrary function $\pi(w)$, the action $do(X = \pi(w))$ represents a model manipulation where the values of a set of variables X are

set to $\pi(w)$ regardless of how the values of X are ordinarily determined in the model via the pre-interventional structural functions. The action $do(X = x)$ where X is set to a constant is the simplest possible intervention and is called atomic. Figs. 3(a,b) show the pre- and post-intervention SCM after action $do(X = x)$ is taken. For simplicity, we'll denote actions $do(X = x), x \in D(X)$ by $do(x)$, so does $P(Y = y|do(X = x)) = P(y|do(x))$.

2.2   Multi-Armed Bandits

We now define the MAB setting using causal language. An agent for a stochastic MAB is given a SCM M with a decision node X representing the arm selection and an outcome variable Y representing the reward – see Fig. 3(a). For arm $x \in D(X)$, its expected reward μ_x is thus the effect of the action $do(x)$, i.e., $\mu_x = \mathbb{E}[Y|do(x)]$. Let μ^* denote the optimal expected reward, $\mu^* = \max_{x \in D(X)} \mu_x$, and x^* the optimal arm. At each trial $t = 1, 2, \ldots, T$, the agent performs an action $do(X_t = x_t)$ and observes a reward Y_t. The objective of the agent is to minimize the cumulative regret, namely,

$$R_T = T\mu^* - \sum_{t=1}^{T} \mathbb{E}[Y_t] = \sum_{x=1}^{K} \Delta_x \mathbb{E}[N_x(T)]$$

We use $KL(\mu_x, \mu^*)$ to denote the Kullback-Leibler divergence between two Bernoulli distributions with mean μ_x and μ^*, and $KL(\mu_x, \mu^*) = \mu_x \log \frac{\mu_x}{\mu^*} + (1 - \mu_x) \log \frac{1-\mu_x}{1-\mu^*}$.

2.3   Identification of Causal Effects

One fundamental problem in causal inference is to estimate the effect of an intervention $do(X = x)$ on an outcome variable Y from a combination of the observational distribution $P(v)$ and causal diagram G (associated with the underlying SCM M). Let $P_M(\cdot)$ denote an arbitrary distribution induced by a SCM M. The following definition captures the requirement that a causal effect be estimable from $P(v)$:

Definition 1 (Identifiability). *[Pearl, 2000, pp. 77] The causal effect of X on Y is identifiable from G if and only if $P_{M_1}(y|do(x)) = P_{M_2}(y|do(x))$ for any pair of models M_1 and M_2 compatible with G such that $P_{M_1}(v) = P_{M_2}(v) > 0$.*

A causal effect $P(y|do(x))$ is non-identifiable if there exists a pair of SCMs generating the same observational distribution ($P(v)$) but a different causal distribution $P(y|do(x))$. The identification of causal effects can be systematically decided using *do-calculus* [Pearl, 1995]. Let X and Z be arbitrary disjoint sets of nodes in a causal diagram G. We denote by $G_{\overline{X}\underline{Z}}$ the subgraph obtained by deleting from G all arrows incoming towards X and all arrows outgoing from Z. The *do-calculus* allows one to see the mapping between observational and experimental distributions whenever certain conditions hold in the graph [Pearl, 2000, pp.85–86].

3   Transfer Learning via Causal Inference

In this section, we connect causal analysis with transfer learning in RL by discussing a transfer scenario between two different MAB agents. We apply do-calculus to systematically estimate the expected reward of each arm for the target agent.

We first consider an off-policy learning problem [Strehl *et al.*, 2010; Swaminathan and Joachims, 2015] between two contextual bandit agents A and A' (Fig. 1(a)). Contextual bandits is a variation of MAB where the agent can observe extra information associated with the reward signal [Langford and Zhang, 2008]. Both agents are equipped with the same sensor and actuator to observe the context U and perform action X. Let ϵ be an independent source of randomness. Agent A follows a policy $do(X = \pi(\epsilon, u))$ [1] and summarizes its experiences as the joint distribution $P(x, y, u)$. Agent A', who is observing A interacting with the environment, wants to be more efficient and reuse the observed $P(x, y, u)$ to find the optimal policy faster. This transfer scenario is summarized as Task I in Fig. 1, where the actions, outcomes, context, and causal structures used by A and A' are identical. Since the optimal action $do(X = x^*)$ for each context $U = u$ can be found by evaluating $x^* = \arg\max_{x \in D(X)} \mathbb{E}[Y|do(x), u]$, the off-policy learning problem is equivalent to identifying the causal effect $\mathbb{E}[Y|do(x), u]$ given the observational distribution $P(x, y, u)$. Indeed, the answer in this case is simply to compute the expected conditional reward given $X = x, U = u$ based on $P(x, y, u)$ and use it as if it were the expected reward. We formally prove this statement through *do-calculus*.

Proposition 1. *For Task I described in Fig. 1, given $P(x, y, u)$, the U-specific causal effect can be written as*

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

Proof. Since $(Y \perp\!\!\!\perp X|U)_{G_{\underline{X}}}$, $P(y|do(x), u) = P(y|x, u)$ by the second rule of *do-calculus*, so does Eqn. 1 holds. $\square$

Proposition 1 provides the formal justification for standard off-policy learning techniques (e.g., propensity score) that use the samples from A and A' interchangeably, often assuming, implicitly, that source and target agents are identical (but for the policy itself). We next consider a more challenging scenario involving the transfer from a contextual bandit agent A (Fig. 1(a)) and a standard MAB agent B (Fig. 1(b)). Agent B has the same actuator as A, but is not equipped with any sensor, thus unable to observe the context U. Fig. 1 summarizes this transfer setting. We want to find the optimal arm x^* for agent B given Agent A's experiences summarized as $P(x, y, u)$. Since $x^* = \arg\max_{x \in D(X)} \mathbb{E}[Y|do(x)]$, this transfer problem is equivalent to identifying the causal effect $\mathbb{E}[Y|do(x)]$.

To obtain the identification formula for this causal effect, we repeatedly apply the rules of do-calculus.

Proposition 2. *For Task II described in Fig. 1, given $P(x, y, u)$, $\mathbb{E}[Y|do(x)]$ equals to*

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

Proof. By basic probabilistic operations,

$$\mathbb{E}[Y|do(x)] = \sum_{u \in D(U)} \mathbb{E}[Y|do(x), u]P(u|do(x))$$

By Prop. 1, $\mathbb{E}[Y|do(x), u] = \mathbb{E}[Y|x, u]$, and $P(u|do(x)) = P(u)$ by the third rule of *do-calculus* since $(U \perp\!\!\!\perp X)_{G_{\overline{x}}}$. $\square$

---

[1] This operation is called *stochastic policy* in the causal literature.
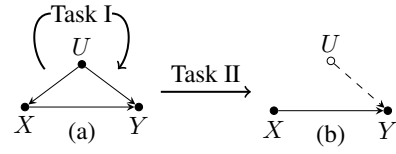

Figure 1: SCMs for (a) a contextual bandit agent. (b) a standard MAB agent.

Prop. 2 provides a mapping between causal effects and the observational distribution. Eq. 2 does not contain any $do(\cdot)$ operator and can be directly estimated from $P(x, y, u)$ – the average effect $\mathbb{E}[Y|do(x)]$ is then said to be identifiable.

The above example shows that transfer learning task can be seen as a problem of identifying causal effects given heterogeneous data (e.g., observational, experimental). Indeed, the do-calculus can be applied to any SCM to identify causal effects using the procedure given in [Tian and Pearl, 2002].

The Challenges of Non-identifiable Tasks

While do-calculus provides a complete method for the identification of causal effects, it cannot construct an identification formula for queries that are not identifiable. To witness, consider a more challenging transfer setting involving the same contextual bandit agent A and a standard MAB agent B as discussed above. Instead of receiving the experiences of A in the form of $P(x, y, u)$, B now can only learn by observing A interacting in the environment – i.e., seeing A's actions and outcomes. Failing to measure the context U, B can only infer the observational distribution $P(x, y)$. Fig. 3 summarizes this transfer scenario as Task 1.

The natural question here is whether the average effect $\mathbb{E}[Y|do(x)]$ can be identified from $P(x, y)$. The do-calculus fails to identify $\mathbb{E}[Y|do(x)]$ in this scenario, which suggests that the average effect in Task 1 is not identifiable.

Proposition 3. *For Task 1 described in Fig. 3, $\mathbb{E}[Y|do(x)]$ is not identifiable from the observational distribution $P(x, y)$.*

Proof. Let $\oplus$ represents the exclusive-or function. Suppose the SCM M encodes the underlying MAB data-generating model, where $U = (U_1, U_2), X, Y, U_1, U_2 \in \{0, 1\}, X = U_1, Y = X \oplus U_1 \oplus U_2, P(U_1 = 0) = P(U_2 = 0) = 0.1$. Expected rewards for arm 0 and 1 are respectively $\mu_0 = 0.18, \mu_1 = 0.82$. Computing $P(x, y)$ from M leads to $\mathbb{E}[Y|X = 0] = \mathbb{E}[Y|X = 1] = 0.9$. Another MAB M' can be constructed such that $U = (U_1, U_2), X, Y, U_1, U_2 \in \{0, 1\}, X = U_1, Y = U_2, P(U_1 = 0) = P(U_2 = 0) = 0.1$. 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. 1). $\square$

Prop. 3 says that $\mathbb{E}[Y|do(x)]$ for the target agent cannot be uniquely computed given prior experiences $P(x, y)$. If one is naive about identifiability and tries to transfer $\mathbb{E}[Y|x]$ as if it were $\mathbb{E}[Y|do(x)]$, negative transfer may occur – i.e., the transferred knowledge will have a negative impact on the performance of the target agent. Unfortunately, in practice, researchers never have access to the underlying SCM and
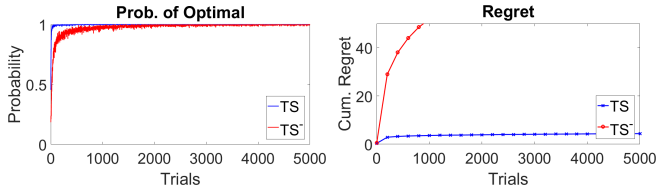
Figure 2: Simulation results of negative transfer example in Sec. 3 comparing the standard Thompson sampling (TS) and Thompson sampling with naive transfer procedure (TS⁻).

therefore cannot distinguish the two models. To illustrate this point, we use the model M provided in the proof of Prop. 3 as the true SCM, and run simulations where 500 samples of $\mathbb{E}[Y|x]$ are naively transferred as if they were $\mathbb{E}[Y|do(x)]$ – see Fig. 2. We can see by inspection the significant disparity between the standard Thompson sampling (TS) [Thompson, 1933; Chapelle and Li, 2011] performance and the Thompson sampling with the naive transfer procedure (TS⁻).

In practice, there exist various transfer settings where the expected reward cannot be identified. We summarize in Figure 3 and Table 1 three canonical tasks where non-identifiability occurs. (C_1, C_2 and C_3 represent the best achievable regret bounds for Tasks 1, 2 and 3, which will be discussed in later sections.) All three tasks represent general settings with a wide range of practical applications. Task 1, as discussed before, models the transfer learning problem between a contextual bandit agent and a standard MAB agent with the context unmeasured (also, when there is a mismatch of context variables). Task 2 and Task 3 describe the transfer learning problem between two agents with different actuators, thus having different action spaces [Argall *et al.*, 2009].

The lack of identifiability in these settings has been understood in the literature. For Task 1, the non-identifiability results from the unobserved condounding between X and Y can be found in [Pearl, 2000, Section 3.5]; non-identifiability from the surrogate $do(z)$ in Task 2 is more subtle, which was shown in [Bareinboim and Pearl, 2012]; [Pearl, 2014] extends this argument and showed that $\mathbb{E}[Y|do(z)]$ cannot be inferred from $P(y|do(x))$ in Task 3.

4 Prior Knowledge as Causal Bounds

One might surmise that the grim results presented so far suggest that when identifiability does not hold, no prior data could be useful and experiments should be conducted from scratch. We will show here that this is not the case. For non-identifiable tasks, we can still obtain bounds over expected rewards of the target agent by constructing a general SCM, in some sense, compatible with all possible models.

We first consider the 2-armed Bernoulli bandits (generalising to higher dimensions emerges naturally) where $X, Y, Z \in \{0, 1\}$. [Pearl, 2000, Section 8.2] constructs a general SCM for Task 2 by projecting $P(u)$ into a prior distribution $P(r_x, r_y)$, where $R_x, R_y \in \{0, 1, 2, 3\}$ and represent all possible functions deciding values of X and Y respectively. We extend this discretization model to bound $\mathbb{E}[Y|do(x)]$ in Task 1. 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
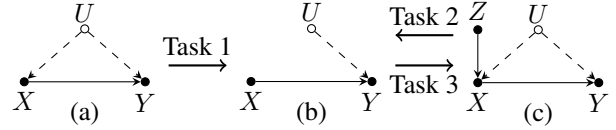

Figure 3: SCM of the three canonical settings where the expected reward is non-identifiable. (a) a contextual bandit agent with context U unmeasured. (b) a standard MAB agent. (c) a standard MAB agent with the action node Z.

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 $\mathcal{O}(\frac{\log(T)}{KL(\mu_x, \mu^*)})$ over the number of draws for a sub-optimal arm x.

$\forall x \in D(X), \forall r_x \in D(R_x), \forall r_y \in D(R_y)$, X and Y are decided by functions $X = f_X(r_x)$ and $Y = f_Y(x, r_y)$ defined as follows:

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

The values of R_y represents all possible function mapping from X to Y and have natural causal interpretation [Heckerman and Shachter, 1995]. Let $p_{ij} = P(X = i, Y = j)$. $P(x, y)$ and $\mathbb{E}[Y|do(x)]$ can then be written as linear combinations in the space spanned by Q:

$$p_{00} = q_{00} + q_{01} \qquad p_{01} = q_{02} + q_{03}$$
$$p_{10} = q_{10} + q_{12} \qquad p_{11} = q_{11} + q_{13} \quad (4)$$

$$\mathbb{E}[Y|do(X = 0)] = q_{02} + q_{03} + q_{12} + q_{13} \quad (5)$$
$$\mathbb{E}[Y|do(X = 1)] = q_{01} + q_{03} + q_{11} + q_{13} \quad (6)$$

We can 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.

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

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

All proofs for theorems are provided in the extended technical report [Zhang and Bareinboim, 2017]. From a causal perspective, this simple bound is unexpected since it was not believed that a model without independence relations (including latents) or exclusion restrictions could impose any informative constraint over the experimental distribution.

We can use similar procedure to discretize the model of Task 3 and decompose U into a 3-tuple (R_z, R_x, R_y), where $R_z \in \{0, 1\}, R_x, R_y \in \{0, 1, 2, 3\}$. We can write

---

Algorithm 1: B-kl-UCB

1: **Input:** A non-decreasing function $f : \mathbb{N} \to \mathbb{R}$
2: A list of bounds over μ_x: $\{[l_x, h_x]\}_{x \in \{1,...,K\}}$
3: **Initialization:** Remove any arm a with $h_x < l_{max}$.
4: Let K' denote the number of remaining arms.
5: Pull each arm of $\{1, \ldots, K'\}$ once
6: **for all** $t = K'$ to $T - 1$ **do**
7:     For each arm x, compute $\hat{U}_x(t) = \min\{U_x(t), h_x\}$, where

$$U_x(t) = \sup\left\{\mu \in [0,1] : KL(\hat{\mu}_x(t), \mu) \le \frac{f(t)}{N_x(t)}\right\}$$

8:     Pick an arm $X_t = \arg\max_{x \in \{1,...,K'\}} \hat{U}_x(t)$.
9: **end for**

Algorithm 2: B-TS for Bernoulli Bandits

1: **Input:** A list of bounds over μ_a: $\{[l_x, h_x]\}_{x \in \{1,...,K\}}$
2: **Initialization:** Remove any arm x with $h_x < l_{max}$.
3: Let K' denote the number of remaining arms.
4: $S_x = 0, F_x = 0, \forall x \in \{1, \ldots, K'\}$
5: **for all** $t = 0$ to $T - 1$ **do**
6:     **for all** $x = 1$ to K' **do**
7:         **repeat**
8:             Draw $\theta_x \sim Beta(S_x + 1, F_x + 1)$
9:         **until** $\theta_x \in [l_x, h_x]$
10:     **end for**
11:     Draw arm $X_t = \arg\max_{x \in \{1,...,K'\}} \theta_x$, observe reward y.
        if $y = 1$ **then** $S_x = S_x + 1$ **else** $F_x = F_x + 1$
12: **end for**

$P(z, y|do(x))$ and $\mathbb{E}[Y|do(z)]$ as a set of linear equations and construct two linear optimization problems that can be solved, and yield a closed-form solution.

Theorem 2. *Consider Task 3 with $X, Y, Z \in \{0, 1\}$, given $P(z, y|do(x))$, for any z, $\mathbb{E}[Y|do(z)] \in [l, h]$, 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. 2 are weaker than Thm. 1 since both arms coincide. [Pearl, 2014] showed that identification of $do(Z)$ is infeasible from experiments over X in task 3, and Thm. 2 provides a stronger condition saying that not even an informative bound can be derived in such settings.

5 Multi-Armed Bandits with Causal Bounds

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

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 [Auer *et al.*, 2002; Cappé *et al.*, 2013]. We extend 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) filtering any arm a during initialization if $h_x < l_{max}$; 2) truncating the UCB $U_x(t)$ with $\hat{U}_x(t) = \min\{U_x(t), h_x\}$ and picking an arm with the largest $\hat{U}_x(t)$. We derive the regret bound for this modification.

Theorem 3. *Consider a K-MAB problem with rewards bounded in $[0, 1]$. For each arm $x \in \{1, \ldots, K\}$ and expected reward μ_x bounded by $[l_x, h_x]$, where $0 < l_x < h_x < 1$.*

Choosing $f(t) = \log(t) + 3\log(\log(t))$, in B-kl-UCB algorithm, the number of draws $\mathbb{E}[N_x(T)]$ for any sub-optimal arm a is upper bounded for any horizon $T \ge 3$ as:

$$\begin{cases} 0 & \text{if } h_x < l_{max} \\ 4 + 4e\log(\log(T)) & \text{if } h_x \in [l_{max}, \mu^*) \\ \frac{\log(T)}{KL(\mu_x, \mu^*)} + \mathcal{O}(\frac{\log(\log(T))}{KL(\mu_x, \mu^*)}) & \text{if } h_x \ge \mu^* \end{cases}$$

Proof. (sketch-see complete proof in the Appendix) Case $h_x < l_{max}$ is obvious. Write $\mathbb{E}[N_x(T)]$ as the sum of two terms: 1) $\sum_{t=K'}^{T-1} \mathbb{P}(\hat{U}_{x^*}(t) < \mu^*)$ and 2) $\sum_{t=K'}^{T-1} \mathbb{P}(\mu^* \le \hat{U}_x(t)), X_t = x)$. Term 1 is upper bounded by $3 + 4e\log(\log(T))$ due to [Cappé *et al.*, 2013, Fact A.1]. Term 2 equals to 0 when $h_x \in [l_{max}, \mu^*)$. When $h_x \ge \mu^*$, Term 2 is upper bounded by $\frac{\log(T)}{KL(\mu_x, \mu^*)} + \mathcal{O}(\frac{\log(\log(T))}{KL(\mu_x, \mu^*)})$ due to [Cappé *et al.*, 2013, Fact A.2]. $\square$

Thm. 3 demonstrates the potential improvements due to the causal bounds. Let $h_x < l_{max}$, $l_{max} \le h_x < \mu^*$ and $h_x \ge \mu^*$ be denoted by cases C_1, C_2, and C_3. If the causal bounds impose strong constraints over the arm's distribution such as cases C_1, C_2, B-kl-UCB provides asymptotic improvements over kl-UCB with bounds over the number of draws for any sub-optimal arm as, respectively, 0 and $\mathcal{O}(\log(\log(T)))$, and dominates kl-UCB. On the other hand, when such constraints are too weak (C_3), the bound of B-kl-UCB degenerates to the standard kl-UCB bound. The proposed algorithm B-kl-UCB, therefore, dominates kl-UCB for all parametrizations. We will show that the latter bound is not improvable for any admissible strategy (not grossly under-performing).

Theorem 4. *(**Lower Bound for** $h_x \ge \mu^*$) Consider a strategy that satisfies $\mathbb{E}[T_x(n)] = o(n^\alpha)$ for any Bernoulli distribution, any arm x with $\Delta_x > 0$, and any $\alpha > 0$. Then, for any arm x with $\Delta_x > 0$ and $h_x > \mu^*$, the following holds*

$$\liminf_{n \to +\infty} \frac{E[N_x(n)]}{\log(n)} \ge \frac{1}{KL(\mu_x, \mu^*)}$$

Remark. These results imply that the constraints imposed by the bounds over the expected rewards (i.e., C_1, C_2, C_3) translate into different regret bounds for the MAB agent. A simple analysis reveals that the three canonical tasks described above can be associated with these different bounds,

(a) Task 1 with parametrizations in C_2 (b) Task 2 with parametrizations in C_1 (c) Task 3 with parametrizations in C_3
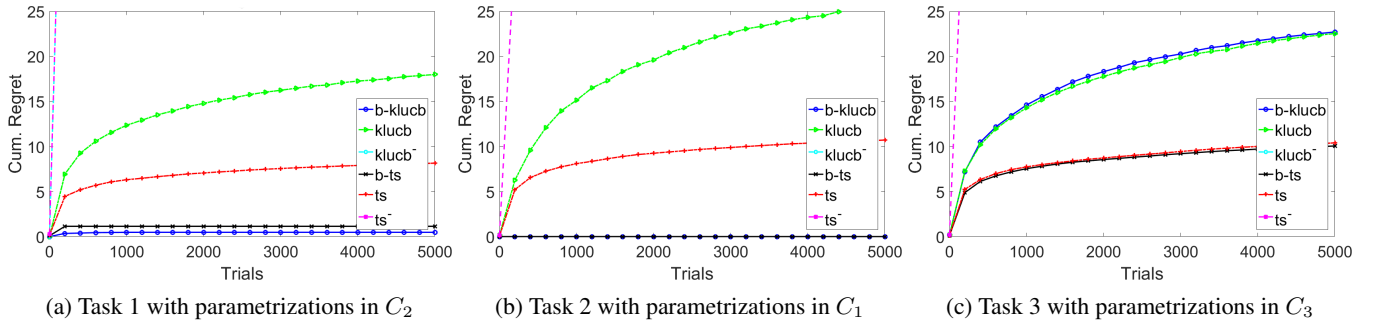
Figure 4: Simulations results of the canonical tasks (Table 1) comparing solvers that are causal enhanced (B-kl-UCB, B-TS), standard (kl-UCB, TS), and naive (kl-UCB$^-$, TS$^-$). Graphs are rendered in high resolution and can be zoomed in.

which is summarized in Table 1. We can see that there exist no parametrization for task 1 satisfying C_1 since $p_{0,1} \leq p_{1,1} + p_{0,0} + p_{0,1}$ and $p_{1,1} \leq p_{0,1} + p_{1,0} + p_{1,1}$ (by Thm. 1). Still, there exists some instances falling into C_2, C_3. Considering the bounds implied by task 2 [Pearl, 2000, pp. 250], we note that there exist a number of instances compatible with C_1 and C_2 (e.g., pick large values for p_{000}, p_{111}), which means that there is great potential for improvement. For task 3, Thm. 2 indicates that the bounds must be the same for both arms, which rules out C_1, C_2 and makes the task to fall into C_3. Since this case is not improvable by Thm. 4, it is the less interesting among the canonical problems.

There exists a class of algorithms based on Thompson sampling (TS) [Thompson, 1933; Chapelle and Li, 2011] that presents strong empirical performance, but its theoretical analysis was not completely understood until recently [Agrawal and Goyal, 2011]. We augment the basic TS solver to take into account the causal bounds, which we call B-TS (Algorithm 2). We employ a rejection sampling approach to enforce the causal bound $[l_x, h_x]$ on the estimated expected reward Θ_x. Simulations 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 open problem.

6 Experimental Results

In this section, we conduct experiments to validate our findings. In particular, we compare B-kl-UCB and B-TS with standard MAB algorithms (kl-UCB and TS) without access to the causal bounds. We also include the counterparts that incorporate a naive transfer procedure described in Sec. 3 (without distinguishing the do-distribution), which we call kl-UCB$^-$ and TS$^-$. We present simulation results for 2-armed Bernoulli bandits. 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 joint distribution. The causal bounds are estimated with the methods described in Sec. 4 from the empirical joint distributions. We assess each algorithm's performance with 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 C_2. The results (Fig. 4a) 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 strategies.

Task 2. The expected rewards of the given param. 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. 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 due to negative transfer.

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 (see Fig. 4c) reveal minor difference 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). kl-UCB$^-$ (CR = 999.8) and TS$^-$ (CR = 1000.00) perform worst due to negative transfer.

These results corroborate with our findings and show that prior experiences can be transferred to improve the performance of the target agent, even when identifiability does not hold. B-kl-UCB dominates kl-UCB in C_1, C_2 while obtains similar performance in C_3. Interestingly, B-TS exhibits similar behaviors as B-kl-UCB in C_1, C_2, and superior performance (with TS) in C_3. This suggests that B-TS may be an attractive practical choice when causal bounds are available.

7 Conclusion

We introduced new learning tools to study settings where the assumption that sensors (contexts) or actuators of the source and target agents perfectly coincide does not hold. In particular, we analyzed the problem of transfer learning across MAB agents in settings where neither do-calculus nor standard off-policy learning techniques can be used due to unobserved confounding. We showed how partial information can still be extracted in these non-identifiable cases, and then translated into potentially informative causal bounds. We incorporated, in a principled way, these bounds into a dynamic allocation procedure and proved regret bounds showing that our algorithm can perform orders of magnitude more efficiently than current, non-causal state-of-the-art procedures.

References

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

[Argall *et al.*, 2009] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

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

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

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

[Bareinboim *et al.*, 2015] Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2015.

[Cappé *et al.*, 2013] 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.

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

[Heckerman and Shachter, 1995] 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.

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

[Konidaris and Barto, 2007] George Konidaris and Andrew G Barto. Building portable options: Skill transfer in reinforcement learning. In *IJCAI*, volume 7, pages 895–900, 2007.

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

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

[Liu and Stone, 2006] Yaxin Liu and Peter Stone. Value-function-based transfer for reinforcement learning using structure mapping. 2006.

[Mehta *et al.*, 2008] Neville Mehta, Soumya Ray, Prasad Tadepalli, and Thomas Dietterich. Automatic discovery and transfer of maxq hierarchies. In *Proceedings of the 25th international conference on Machine learning*, pages 648–655. ACM, 2008.

[Mehta *et al.*, 2011] Neville Mehta, Soumya Ray, Prasad Tadepalli, and Thomas Dietterich. Automatic discovery and transfer of task hierarchies in reinforcement learning. *AI Magazine*, 32(1):35–51, 2011.

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

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

[Pearl, 2014] Judea 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.

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

[Strehl *et al.*, 2010] Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, pages 2217–2225, 2010.

[Swaminathan and Joachims, 2015] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*, pages 814–823, 2015.

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

[Thompson, 1933] 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.

[Tian and Pearl, 2002] J. Tian and J. Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 567–573, Menlo Park, CA, 2002. AAAI Press/The MIT Press.

[Zhang and Bareinboim, 2017] Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandit: a causal approach. Technical Report R-25, Purdue AI Lab., 2017.