Transfer Learning in Multi-Armed Bandit: A Causal Approach

(Extended Abstract)

Junzhe Zhang
Purdue University
zhang745@purdue.edu

Elias Bareinboim
Purdue University
eb@purdue.edu

ABSTRACT

We leverage causal inference tools to support a principled and more robust transfer of knowledge in reinforcement learning (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 nor standard off-policy 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 novel bandit algorithm, B-kl-UCB. Simulations 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

Reinforcement learning (RL) agents are typically trained 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 [13, 10, 12, 11, 17]. Causal inference deals with the problem of inferring the effect of actions (target) from a combination of a structural model (to be defined) and heterogeneous sources of data while permitting the presence of unmeasured common causes (also called unobserved confounders, or UCs) [15, 5]. In his seminal work, [14] developed a general calculus known as *do-calculus* which was shown to be complete for observational and experimental identification, i.e., any causal effect can be identified from an observational or experimental dataset if and only if it can be derived by do-calculus [18, 16, 8, 4]. Connections between causal models with UCs and RL were first established in [3]. Nevertheless, their methods mainly focused on the online learning scenarios and barely touched transfer learning settings.

This paper considers 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. For three canonical tasks where the causal effect is not identifiable, we extract knowledge from the available distributions as bounds over the expected reward (called *causal bounds*). We propose a novel MAB algorithm (B-kl-UCB) that takes these bounds as input and empirically show that the regret bound of B-kl-UCB dominates the standard kl-UCB [6].

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

2. TRANSFER IN CAUSAL SEMANTICS

We define the MAB setting using structural causal models (SCMs) [15, Sec. 3], which serve as the basic semantical framework of our analysis. A causal diagram associated with the SCM M is a directed acyclic graph where solid nodes correspond to observed variables, empty nodes correspond to unobserved variables, and edges represent functional relationships. 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. 1(b). We use the $do(\cdot)$ operator to denote interventions (actions) [15, Sec. 3]. Let $D(X)$ denote by the domain of variable X. For arm $x \in D(X)$, its expected reward μ_x is thus the effect of the action $do(X = x)$, i.e., $\mu_x = \mathbb{E}[Y|do(X = x)]$. Let μ^* denote the optimal expected reward, $\mu^* = \max_{x \in D(X)} \mu_x$. 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 $R_T = T\mu^* - \sum_{t=1}^{T} \mathbb{E}[Y_t]$. We will consistently use the abbreviation $P(x)$ for the probabilities $P(X = x)$ and $do(x)$ for actions $do(X = x)$.

A transfer learning problem between two bandit agents can thus be defined as the identification of the expected reward of the target agent (e.g., $\mathbb{E}[Y|do(x)]$) given the causal diagram and trajectories of the source agent (e.g., $P(x, y)$). We summarize in Figure 1 and Table 1 three canonical TL tasks where target causal effects are not identifiable [1], that is, the *do-calculus* is unable to pin down the mapping between the target causal effect and the source distribution [15, pp. 77]. All three tasks are practical problems which

[1]For identifiable tasks, causal effects can be estimated by repeatedly applying the *do-calculus*.

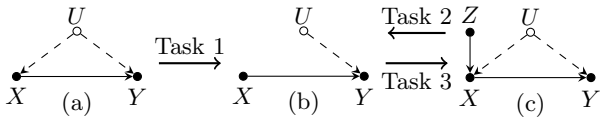

Figure 1: SCM of the three canonical settings where the expected reward is non-identifiable. (a) a contextual bandit agent with context U unmeasured. (b) a standard MAB agent. (c) a standard MAB agent with the action node Z.

Task	Source → Target		ID
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: Three canonical transfer learning tasks. ID stands for point identifiability.

have a broad range of practical applications. Task 1 models the transfer learning problem between a contextual bandit agent and a standard MAB agent with the context U unmeasured. Tasks 2 and Task 3 describe the transfer learning problem between two agents with different actuators, thus having different action spaces [1].

3. MABS WITH CAUSAL BOUNDS

One might surmise that the negative 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 obtain bounds over expected rewards of the target agent and propose a novel MAB algorithm leveraging these causal bounds.

Consider the 2-armed Bernoulli bandits (generalizing to higher dimensions emerges naturally) where $X, Y, Z \in \{0, 1\}$. Construct a discretized SCM for Task 2 by decomposing U into a pair of canonical types (R_x, R_y) [15, Sec. 8.2], where $R_x, R_y \in \{0, 1, 2, 3\}$ and represent the different types of individuals in the population [9, 7]. We now extend this discretization model to bound $\mathbb{E}[Y|do(x)]$ in Task 1 given $P(x, y)$. Let $q_{ij} = P(R_x = i, R_y = j) \geq 0$, and $Q = \{q_{ij}\}$. $P(x, y)$ and $\mathbb{E}[Y|do(x)]$ can then be written as linear combinations in the space spanned by Q. We then obtain causal bounds by optimizing (min or max) $\mathbb{E}[Y|do(x)]$ subject to constraints $P(x, y)$ and $q_{ij} \geq 0$. The similar procedure can also be applied to bound $\mathbb{E}[Y|do(z)]$ in Task 3.

We now consider how the causal bounds can be used to efficiently identify an optimal arm. We extend UCB algorithms [2, 6] to take into account the causal bounds, which we call B-kl-UCB (Algorithm 1). Let l_{max} denote the maximum of all lower bounds by $l_{max} = \max_{x=1,\dots,K} l_x$. 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 $\hat{U}_x(t)$.

4. EXPERIMENTAL RESULTS

We conduct experiments for Task 1 with 2-armed Bernoulli bandits. We compare B-kl-UCB with the standard kl-UCB

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: $\{[l_x, h_x]\}_{x\in\{1,\dots,K\}}$
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 $\{1, \dots, K'\}$ 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

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

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

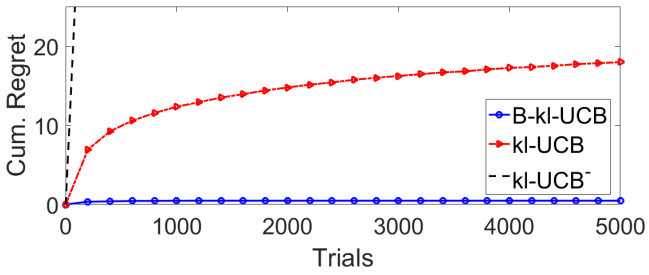

Figure 2: Simulations results of Task 1 (Table 1) comparing solvers that are causal enhanced (B-kl-UCB), standard (kl-UCB), and naive (kl-UCB$^-$).

algorithm without access to the causal bounds. We also include its counterpart (called kl-UCB$^-$) that incorporates a naive transfer procedure using observational expected reward $\mathbb{E}[Y|x]$ as if it was the average causal effect $\mathbb{E}[Y|do(x)]$. 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 methods described in Sec. 3 from the corresponding empirical joint distributions. We assess each algorithm's performance with cumulative regrets (CR). The expected rewards of the given parametrization are $\mu_1 = 0.66, \mu_2 = 0.36$, and the estimated causal bounds are $b_1 = [0.03, 0.76], b_2 = [0.21, 0.51]$. The results (Fig. 2) reveal a significant difference in the regret experienced by B-kl-UCB (CR = 0.47) compared to kl-UCB (CR = 17.97). kl-UCB$^-$ performs worst among all strategies (CR = 1499.70). These results corroborate with our methods and show that prior experiences can be transferred to improve the performance of the target agent, even when identifiability does not hold.

5. CONCLUSION

We tackled the problem of transfer learning across MAB agents in general canonical settings where neither do-calculus nor standard 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 these bounds into a dynamic allocation procedure and empirically showed that our algorithm can perform orders of magnitude more efficiently than current, non-causal state-of-the-art procedures.

REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] G. W. Imbens and D. B. Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, pages 305–327, 1997.

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

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

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

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

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

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

[16] 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.

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

[18] 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.