Learning Endogenous Representation in Reinforcement Learning via Advantage Estimation

Hsiao-Ru Pan **Bernhard Schölkopf**
Max Planck Institute for Intelligent Systems, Tübingen
{hpan,bs}@tuebingen.mpg.de

Abstract

Recently, it was shown that the advantage function can be understood as quantifying the causal effect of an action on the cumulative reward. However, this connection remained largely analogical, with unclear implications. In the present work, we examine this analogy using the Exogenous Markov Decision Process (ExoMDP) framework, which factorizes an MDP into variables that are causally related to the agent's actions (endogenous) and variables that are beyond the agent's control (exogenous). We demonstrate that the advantage function can be expressed using only the endogenous variables, which is, in general, not possible for the (action-)value function. Through experiments in a toy ExoMDP, we found that estimating the advantage function directly can facilitate learning representations that are invariant to the exogenous variables.

1 Introduction

Imagine playing the game of chess in an outdoor environment, where the lighting conditions are subject to change due to the time of the day, the weather, or maybe the shadow of an object passing by. While these *exogenous* conditions can change the way we perceive the chessboard, an experienced player should be able to quickly recognize the *endogenous* state of the game irrespective of the *exogenous* conditions.

As the field of machine learning is moving towards increasingly high-dimensional and complex problems, learning representations that capture the causal relationships between observations and downstream tasks can be crucial to generalization or robustness [Schölkopf et al., 2021].

Recently, Pan et al. [2022] demonstrated that the advantage function in reinforcement learning (RL) can be understood as quantifying the causal effect of actions on the cumulative reward through an analogy to the treatment effect in the Neyman-Rubin Causal Model [Splawa-Neyman et al., 1990, Rubin, 1974]. We shall examine this analogy more closely using the Exogenous Markov Decision Process [Efroni et al., 2022] framework, which gives precise definitions of endogenous and exogenous variables.

Our contributions can be summarized as follows:

- We show that the advantage function is endogenous in nature in the sense that it can be expressed using only the endogenous variables of an ExoMDP.
- We demonstrate empirically that Direct Advantage Estimation can facilitate learning representations that are invariant to exogenous variables.

Accepted to the NeurIPS 2023 Workshop on Causal Representation Learning.

2 Background

A Markov Decision Process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, p, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the transition probability (commonly written as $p(s'|s, a)$), and $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function [Sutton et al., 1998]. For simplicity, we will consider MDPs that reach absorbing states with probability 1. The following analysis can also be trivially extended to discounted infinite horizon problems. The goal of RL is to learn a policy, a mapping from states to distributions over actions, that maximizes the expected cumulative reward. More formally, we seek $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ (Δ denotes distributions over the given set) such that $J(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} r(s_t, a_t)]$ is maximized. Typically, this is achieved by iteratively estimating the expected return given the policy ($Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} r_t|s_0{=}s, a_0{=}a]$ or $V^\pi(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} r_t|s_0{=}s]$) and improving the policy based on the estimated return.

2.1 Exogenous MDP

To give a more detailed view of the data-generating process, we adopt the Exogenous MDP (ExoMDP) framework [Efroni et al., 2022]. In an ExoMDP, the state space is factorized into two components $\mathcal{S} = \mathcal{S}_e \times \mathcal{S}_x$, namely the **e**ndogenous component and the e**x**ogenous component. These two components differ in how they depend on the actions, more specifically, the transition probability is factorized into $p(s_{t+1}|s_t, a_t) = p_e(s^e_{t+1}|s^e_t, a_t)p_x(s^x_{t+1}|s^x_t)$, meaning that only the transitions of the endogenous component depend on the actions but not the exogenous component. Similarly, the reward function can be decomposed into $r(s, a) = r_e(s^e, a) + r_x(s^x)$[1]. See Figure 1 for a comparison between the graphical models of MDPs and ExoMDPs. One consequence of this factorization is that, if the factorization of a given ExoMDP is known a priori, then it can be solved by solely considering the endogenous part of the MDP, which can drastically reduce the complexity of the problem when $|\mathcal{S}_{\text{end}}| \ll |\mathcal{S}_{\text{exo}}|$.

One interesting aspect of ExoMDP is that it enables us to quantitatively determine how much of the rewards are *caused* by the agent's actions. More specifically, from Figure 1, it can be seen that the exogenous rewards are not caused by the actions as there are no chains connecting these variables.

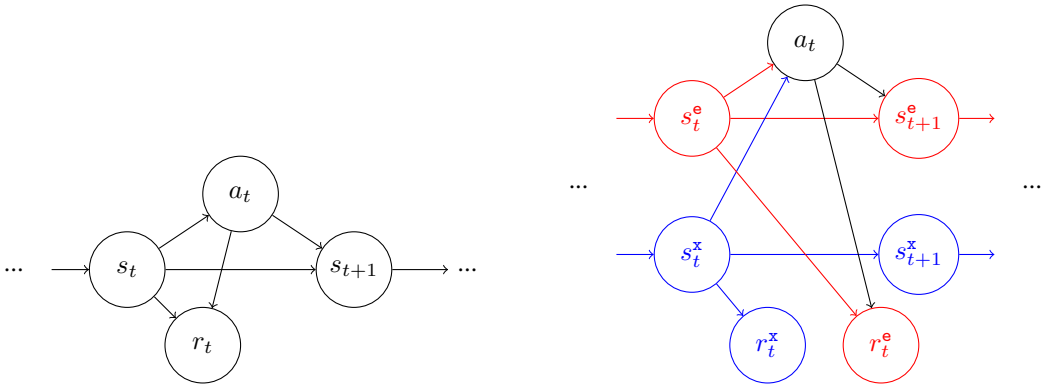


Figure 1: Graphical models of MDPs (left) and ExoMDPs (right). Colors and superscripts indicate whether the variable or relationship is **e**ndogenous (red) or e**x**ogenous (blue).

2.2 Direct Advantage Estimation (DAE)

Aside from Q^π and V^π, another function of interest in RL is the advantage function, defined by $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$ Baird [1994]. It was shown that the advantage function can reduce the variance of policy optimization methods and has been widely used in practice [Greensmith et al., 2004, Kakade and Langford, 2002]. More recently, Pan et al. [2022] showed that the advantage function can be understood as the causal effect of an action on the return through an analogy with the treatment effect from the Neyman-Rubin causal model [Splawa-Neyman et al., 1990, Rubin, 1974].

[1] In the original definition, only r_e is used. Here, we use a more general definition by allowing an exogenous reward function r_x.

This analogy relies on the fact that both the treatment effect and the advantage function compare the outcome of taking a certain treatment (action) to what would have happened otherwise. However, the connection between the causal effect and the advantage function remains largely an analogy, and the extent to which it holds remains unanswered.

In addition to the analogy, they also showed that the advantage function can be estimated directly by minimizing the following constrained objective:

$$V^\pi, A^\pi = \arg\min_{\hat{V},\hat{A}} \mathbb{E}\left[\left(\sum_{t=0}^{\infty}\Big(r(s_t,a_t) - \hat{A}(s_t,a_t)\Big) - \hat{V}(s_0)\right)^2\right] \tag{1}$$

$$\text{subject to} \sum_{a\in\mathcal{A}} \hat{A}(s,a)\pi(a|s) = 0. \tag{2}$$

It was reported that estimating the advantage function this way can lead to superior policy optimization performance in various domains compared to previous methods.

3 Endogeneity of the Advantage Function

To demonstrate the endogenous nature of the advantage function, we show that: (1) The advantage function is invariant to exogenous rewards, and (2) If the policy is endogenous, i.e., it only depends on the endogenous variables, then the advantage function can be expressed solely using the endogenous part of the ExoMDP.

Lemma 1. Given an ExoMDP $(\mathcal{S} = \mathcal{S}_e \times \mathcal{S}_x, \mathcal{A}, p = p_e p_x, r = r_e + r_x)$, and a policy π, then the following holds:

$$A^\pi(s,a) = \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{\infty} r_e(s_t^e,a_t)\,\middle|\, s_0{=}s, a_0{=}a\right] - \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{\infty} r_e(s_t^e,a_t)\,\middle|\, s_0{=}s\right], \tag{3}$$

where s_t^e is the endogenous state at time t.

Proof. By definition,

$$A^\pi(s,a) = \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{\infty} r(s_t,a_t)\,\middle|\, s_0{=}s, a_0{=}a\right] - \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{\infty} r(s_t,a_t)\,\middle|\, s_0{=}s\right] \tag{4}$$

$$= \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{\infty} r_e(s_t^e,a_t) + r_x(s_t^x)\,\middle|\, s_0{=}s, a_0{=}a\right] - \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{\infty} r_e(s_t^e,a_t) + r_x(s_t^x)\,\middle|\, s_0{=}s\right] \tag{5}$$

$$= \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{\infty} r_e(s_t^e,a_t)\,\middle|\, s_0{=}s, a_0{=}a\right] - \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{\infty} r_e(s_t^e,a_t)\,\middle|\, s_0{=}s\right]. \tag{6}$$

The last equation holds, as r_x does not depend on the action. This shows that the advantage function only depends on the endogenous rewards, which concludes the proof. □

For example, shifting the reward function by a constant (i.e., $\tilde{r}(s,a) = r(s,a) + c$), has no effect on the advantage function, as constant rewards do not depend on actions. For the second property,

Theorem 1. Following the notations in Lemma 1, if there exists $\pi_e : \mathcal{S}_e \to \Delta(\mathcal{A})$ such that $\pi(a|s = (s^e, s^x)) = \pi_e(a|s^e)$, then there exists $A_e^{\pi_e} : \mathcal{S}_e \times \mathcal{A} \to \mathbb{R}$ such that

$$A^\pi(s,a) = A_e^{\pi_e}(s^e,a) \tag{7}$$

for all $s = (s^e, s^x) \in \mathcal{S}$, $a \in \mathcal{A}$.

Proof. Since $\pi(a|s) = \pi_e(a|s^e)$, we have

$$p(s_{t+1}^e|s_t) = \sum_{a\in\mathcal{A}} p(s_{t+1}^e|s_t,a)\pi(a|s_t) = \sum_{a\in\mathcal{A}} p(s_{t+1}^e|s_t^e,a)\pi_e(a|s_t^e) = p(s_{t+1}^e|s_t^e). \tag{8}$$

This means that the distribution of the endogenous states no longer depends on the exogenous states. This can also be seen from the causal graph in Figure 1 by breaking the arrows from s_t^x to a_t. Through a similar argument, we have $p(s_{t'}^e|s_t) = p(s_{t'}^e|s_t^e)$ for any $t' > t$. Following the result from Lemma 1, we have

$$A^\pi(s,a) = \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{\infty} r_e(s_t^e, a_t)\middle| s_0{=}s, a_0{=}a\right] - \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{\infty} r_e(s_t^e, a_t)\middle| s_0{=}s\right] \tag{9}$$

$$= \mathop{\mathbb{E}}_{\pi_e}\left[\sum_{t=0}^{\infty} r_e(s_t^e, a_t)\middle| s_0^e{=}s^e, a_0{=}a\right] - \mathop{\mathbb{E}}_{\pi_e}\left[\sum_{t=0}^{\infty} r_e(s_t^e, a_t)\middle| s_0^e{=}s^e\right] =: A_e^{\pi_e}(s^e, a). \tag{10}$$

□

We note that, in general, neither Q^π nor V^π satisfy this kind of equivalence, as exogenous rewards are also accumulated in the expectation[2]. However, by taking the difference of the two, we essentially remove the exogenous component, as in the case of the advantage function. This property suggests that a representation of the state that depends solely on the endogenous part of the ExoMDP is enough to estimate the advantage function. An interesting question to ask is, then, whether this holds in the opposite direction, i.e., does estimating the advantage function lead to representations that are endogenous? In the next section, we shall examine this question empirically using a toy ExoMDP.

4 Experiments

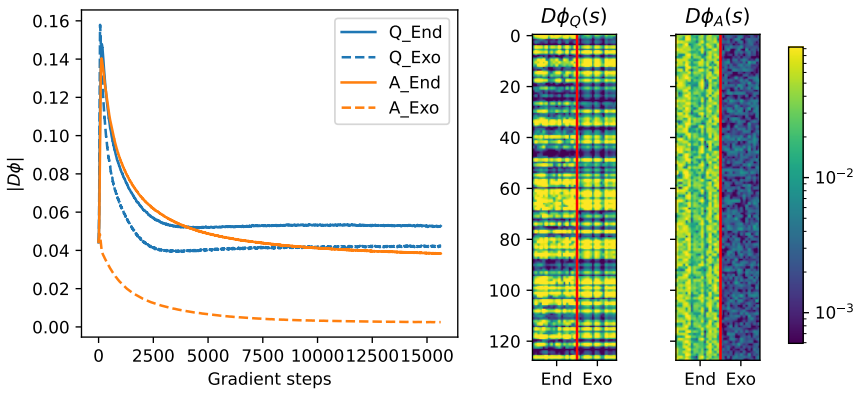
In this section, we compare the representations learned by estimating the Q-function with Monte Carlo methods, and the advantage function with DAE using neural networks.

Environment We use a toy linear finite-horizon ExoMDP with $\mathcal{S} = \mathcal{S}_e \oplus \mathcal{S}_x = \mathbb{R}^{16} \oplus \mathbb{R}^{16}$, discrete action space $\mathcal{A} = \{0, 1, 2, 3\}$, deterministic linear transitions with $s' = (T_{e,a} \oplus T_x)(s)$ ($T_{e,a}$ and T_x are 16×16 matrices), reward functions $r(s,a) = (r_{e,a} \oplus r_x)(s)$ ($r_{e,a}$ and r_x are linear functionals), and horizon $H = 16$. To assess whether the learned representation can disentangle endogenous variables from exogenous variables, we use a mixing function M such that the agent does not directly observe the underlying state variables but only $M(s)$. Here, we consider three different mixing functions: (1) Identity (no mixing), (2) Linear transformation (a random matrix), and (3) Nonlinear transformation (a randomly initialized neural network with one hidden layer).

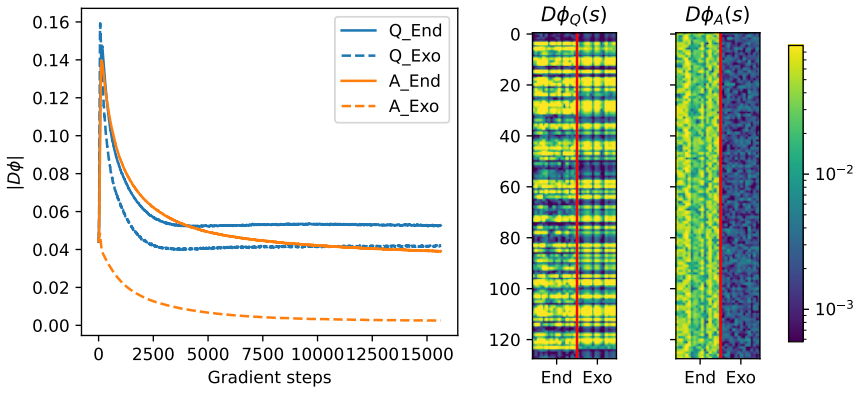
Agent We use a neural network with two hidden layers of size 128 with tanh activations and an output layer of size $|\mathcal{A}| = 4$ to approximate Q^π or A^π. The output of the last hidden layer, denoted $\phi(s) \in \mathbb{R}^{128}$, is referred to as the learned representation. The agents are trained online with trajectories sampled using the uniform policy. For each gradient step, we sample 64 trajectories and compute the gradient with respect to the objective functions (Equation 1 or $(\hat{Q}(s_{t'}, a_{t'}) - \sum_{t=t'}^{H-1} r_t)^2$ in the case of learning the Q-function). In sum, we perform 15625 gradient steps, using $64 \times 15625 = 10^6$ sampled trajectories.

Results We measure how much the learned representation depends on the state variables through the Jacobian matrix $D\phi(s)$. The results are summarized in Figure 2. From the Jacobian matrices, we see that learning Q^π and A^π can lead to drastically different representations. In the case of ϕ_Q, we found both endogenous and exogenous variables to have similar levels of influence on the representation, with a slightly stronger dependence on the endogenous variables. On the other hand, we see $D\phi_A$ behaves very differently for endogenous and exogenous variables, where the entries are close to zero for the exogenous variables while the entries for the endogenous variables remain comparable to $D\phi_Q$. This suggests that the representations learned by DAE tend to be more *causal* in the sense that they capture the dependencies on the endogenous variables while being more invariant to exogenous ones. Perhaps more surprising are the cases when mixing of state variables is involved, we see DAE is still capable of learning representations that have weak dependencies on the exogenous variables.

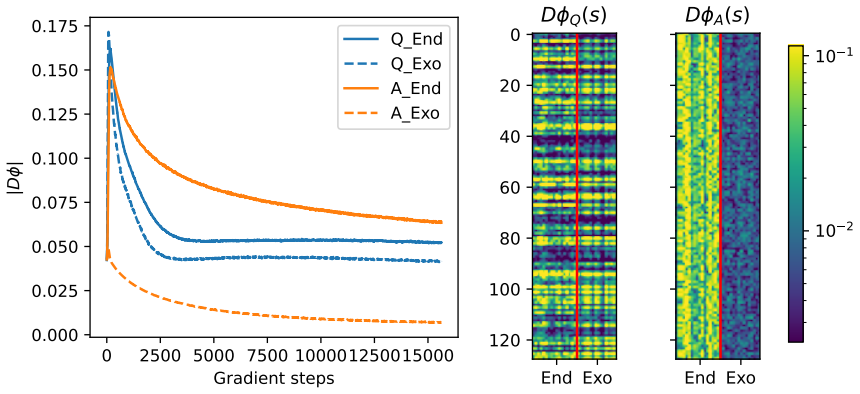
[2] In the original work by Efroni et al. [2022], it was assumed that all rewards are endogenous, which makes it possible to express both Q^π and V^π using only the endogenous variables.



(a) Identity



(b) Linear



(c) Nonlinear

Figure 2: Comparison of the Jacobian matrices of the learned representations under different mixing functions and training objectives. **Left**: Average of element-wise absolute value of $D\phi(s)$ over endogenous/exogenous variables (element-wise) and sampled trajectories throughout training. Lines and dashes indicate endogenous and exogenous variables, respectively. Colors indicate the objective Q^π (blue) and A^π (orange). Curves are averaged over 25 runs. **Right**: The element-wise absolute value of the Jacobian matrices (averaged over 64 trajectories). The Y-axes label the dimension of the representation. Subscripts denote the learning objective.

5 Related work

Causality is ubiquitous in RL, as the interactive nature of RL requires the agent to estimate the effects of its actions on the environment. One major line of work in the intersection of causality and RL lies in the study of decision-making in the presence of unobserved confounders, which dates back at least to Splawa-Neyman et al. [1990], Rubin [1974]. More recently, similar problems have been studied in the bandit setting [Bareinboim et al., 2015, Sen et al., 2017], or the sequential setting [Tennenholtz et al., 2020, Zhang and Bareinboim, 2016]. Another line of work, to which our work is more related, has focused on learning representations that encode the causal relationship between states and downstream tasks [Zhang et al., 2020b,a, Trimponias and Dietterich, 2023].

6 Discussion

In the present work, we have demonstrated that the advantage function is endogenous in nature, a property that is not shared by the (action-)value function. This property led us to hypothesize that estimating the advantage function directly can lead to representations that are also endogenous. Through experiments in a simulated environment, we found that it indeed facilitates learning representations that are endogenous. Finally, we note some limitations of this work. (1) While we demonstrated that learning the advantage function via DAE can lead to endogenous representations, a theoretical justification for why this is the case remains desirable. (2) One motivation to factorize an MDP into an ExoMDP is to reduce the effective problem size and improve sample efficiency; however, it is not clear how the learned representations can be utilized for more efficient deep RL algorithms.

References

L. C. Baird. Reinforcement learning in continuous time: Advantage updating. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 4, pages 2448–2453. IEEE, 1994.

E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28, 2015.

Y. Efroni, D. J. Foster, D. Misra, A. Krishnamurthy, and J. Langford. Sample-efficient reinforcement learning in the presence of exogenous information. In *Conference on Learning Theory*, pages 5062–5127. PMLR, 2022.

E. Greensmith, P. L. Bartlett, and J. Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(9), 2004.

S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

H.-R. Pan, N. Gürtler, A. Neitz, and B. Schölkopf. Direct advantage estimation. *Advances in Neural Information Processing Systems*, 35:11869–11880, 2022.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

R. Sen, K. Shanmugam, M. Kocaoglu, A. Dimakis, and S. Shakkottai. Contextual bandits with latent confounders: An nmf approach. In *Artificial Intelligence and Statistics*, pages 518–527. PMLR, 2017.

J. Splawa-Neyman, D. M. Dabrowska, and T. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

R. S. Sutton, A. G. Barto, et al. Introduction to reinforcement learning. 1998.

G. Tennenholtz, U. Shalit, and S. Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10276–10283, 2020.

G. Trimponias and T. G. Dietterich. Reinforcement learning with exogenous states and rewards. *arXiv preprint arXiv:2303.12957*, 2023.

A. Zhang, C. Lyle, S. Sodhani, A. Filos, M. Kwiatkowska, J. Pineau, Y. Gal, and D. Precup. Invariant causal prediction for block mdps. In *International Conference on Machine Learning*, pages 11214–11224. PMLR, 2020a.

A. Zhang, R. McAllister, R. Calandra, Y. Gal, and S. Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020b.

J. Zhang and E. Bareinboim. Markov decision processes with unobserved confounders: A causal approach. Technical report, Technical report, Technical Report R-23, Purdue AI Lab, 2016.