Forward and Backward State Abstractions for Off-policy Evaluation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Off-policy evaluation (OPE) is crucial for evaluating a target policy's impact offline before its deployment. However, achieving accurate OPE in large state spaces remains challenging. This paper studies state abstractions – originally designed for policy learning – in the context of OPE. Our contributions are three-fold: (i) We define a set of irrelevance conditions central to learning state abstractions for OPE. (ii) We derive sufficient conditions for achieving irrelevance in Q-functions and marginalized importance sampling ratios, the latter obtained by constructing a time-reversed Markov decision process (MDP) based on the observed MDP. (iii) We propose a novel two-step procedure that sequentially projects the original state space into a smaller space, which substantially simplify the sample complexity of OPE arising from high cardinality.

1 Introduction

Motivation. Off-policy evaluation (OPE) serves as a crucial tool for assessing the impact of a newly developed policy using a pre-collected historical data before its deployment in high-stake applications, such as healthcare (Murphy et al., 2001), recommendation systems (Chapelle & Li, 2011), education (Mandel et al., 2014), dialog systems (Jiang et al., 2021) and robotics (Levine et al., 2020). A fundamental challenge in OPE is its "off-policy" nature, wherein the target policy to be evaluated differs from the behavior policy that generates the offline data. This distributional shift is particularly pronounced in environments with large state spaces of high cardinality. Theoretically, the minimax rate for estimating the target policy's Q-function decreases rapidly as the state space dimension increases (Chen & Qi, 2022). Empirically, large state space significantly challenges the performance of state-of-the-art OPE algorithms (Fu et al., 2020; Voloshin et al., 2021).

Although different policies induce different trajectories in the large ground state space, they can produce similar paths when restricted to relevant, lower-dimensional state spaces (Pavse & Hanna, 2023). Consequently, applying OPE to these abstract spaces can significantly mitigate the distributional shift between target and behavior policies, enhancing the accuracy in predicting the target policy's value. This makes state abstraction, designed to reduce state space cardinality, particularly appealing for OPE. However, despite the extensive literature on studying state abstractions for policy learning (see Section 1.1 for details), it has been hardly explored in the context of OPE.

Contributions. This paper aims to systematically investigate state abstractions for OPE to address the aforementioned gap. Our main contributions include:

1. Introduction of a set of irrelevance conditions for OPE, accompanied by validations of various OPE methods when applied to abstract state spaces under these conditions.
2. Derivation of sufficient conditions for state abstractions to achieve irrevelance in Q-functions and marginalized importance sampling (MIS) ratios. A key ingredient of our proposal lies in

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

constructing a time-reversed Markov decision process (MDP, Puterman, 2014) by swapping the future and past. This effectively yields state abstractions that achieve the irrelevance property.

3. Development of a novel two-step procedure to sequentially obtain a smaller state space and reduce the sample complexity of OPE. It is also guaranteed to yield a smaller state space compared to existing single-step abstractions.

1.1 Related work

Our proposal is closely related to OPE and state abstraction. Additional related work on confounder selection in causal inference is relegated to Appendix A.

Off-policy evaluation. OPE aims to estimate the average return of a given target policy, utilizing historical data generated by a possibly different behavior policy (Dudík et al., 2014; Uehara et al., 2022). The majority of methods in the literature can be classified into the following three categories:

1. **Value-based methods** that estimate the target policy's return by learning either a value function (Sutton et al., 2008; Luckett et al., 2019; Li et al., 2024) or a Q-function (Le et al., 2019; Feng et al., 2020; Hao et al., 2021; Liao et al., 2021; Chen & Qi, 2022; Shi et al., 2022) from the data.
2. **Importance sampling (IS) methods** that adjust the observed rewards using the IS ratio, i.e., the ratio of the target policy over the behavior policy, to address their distributional shift. There are two major types: sequential IS (SIS, Precup, 2000; Thomas et al., 2015; Hanna et al., 2019; Hu & Wager, 2023) which employs a cumulative IS ratio, and marginalized IS (Liu et al., 2018; Nachum et al., 2019; Xie et al., 2019; Dai et al., 2020; Yin & Wang, 2020; Wang et al., 2023) which uses the MIS ratio to mitigate the high variance of the SIS estimator.
3. **Doubly robust methods** or their variants that employ both the IS ratio and the value/reward function to enhance the robustness of OPE (Zhang et al., 2013; Jiang & Li, 2016; Thomas & Brunskill, 2016; Farajtabar et al., 2018; Kallus & Uehara, 2020; Tang et al., 2020; Uehara et al., 2020; Shi et al., 2021; Kallus & Uehara, 2022; Liao et al., 2022; Xie et al., 2023).

However, none of the aforementioned works studied state abstraction, which is our primary focus.

State abstraction. State abstraction aims to obtain a parsimonious state representation to simplify the sample complexity of reinforcement learning (RL), while ensuring that the optimal policy restricted to the abstract state space attains comparable values as in the original, ground state space. There is an extensive literature on the theoretical and methodological development of state abstraction, particularly bisimulation — a type of abstractions that preserve the Markov property in the abstracted state (Singh et al., 1994; Dean & Givan, 1997; Givan et al., 2003; Ravindran, 2004; Jong & Stone, 2005; Li et al., 2006; Ferns et al., 2004, 2011; Pathak et al., 2017; Wang et al., 2017; Ha & Schmidhuber, 2018; François-Lavet et al., 2019; Gelada et al., 2019; Castro, 2020; Zhang et al., 2020; Allen et al., 2021; Abel, 2022). In particular, Li et al. (2006) analyzed five irrelevance conditions for optimal policy learning. Unlike the aforementioned works that focus on policy learning, we introduce irrelevance conditions for OPE, and propose abstractions that satisfy these irrelevant properties. Meanwhile, the proposed abstraction for achieving irrelevance for the MIS ratio resembles the Markov state abstraction developed by Allen et al. (2021) in the context of policy learning.

More recently, Pavse & Hanna (2023) made a pioneering attempt to study state abstraction for OPE, proving its benefits in enhancing OPE accuracy. However, they primarily focused on MIS estimators. In contrast, our theoretical analysis applies to a broader range of estimators. Moreover, their abstraction did not achieve MIS-ratio irrelevance, nor did they implement the two-step procedure.

Lastly, state abstraction is also related to variable selection (Tangkaratt et al., 2016; Wang et al., 2017; Zhang & Zhang, 2018; Ma et al., 2023) and representation learning for RL (Abel et al., 2016; Shelhamer et al., 2016; Laskin et al., 2020; Uehara et al., 2021).

2 Preliminaries

In this section, we first introduce some key concepts relevant to OPE in RL, such as MDP, target and behavior policies, value functions, IS ratios (Section 2.1). We next review state abstractions for optimal policy learning (Section 2.2), alongside with four prominent OPE methodologies (Section 2.3).

2.1 Data generating process, policy, value and IS ratio

Data. Assume the offline dataset $\mathcal{D}$ comprises multiple trajectories, each containing a sequence of state-action-reward triplets $(S_t, A_t, R_t)_{t\geq 1}$ following a finite MDP, denoted by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \rho_0, \gamma \rangle$. Here, $\mathcal{S}$ and $\mathcal{A}$ are the discrete state and action spaces, both with finite cardinalities, $\mathcal{T}$ and $\mathcal{R}$ are the state transition and reward functions, ρ_0 denotes the initial state distribution, and $\gamma \in (0, 1)$ is the discount factor.

The data is generated as follows: (i) At the initial time, the state S_1 is generated according to ρ_0; (ii) Subsequently, at each time t, the agent finds the environment in a specific state $S_t \in \mathcal{S}$ and selects an action $A_t \in \mathcal{A}$ according to a behavior policy b such that $\mathbb{P}(A_t = a|S_t) = b(a|S_t)$; (iii) The environment delivers an immediate reward R_t with an expected value of $\mathcal{R}(A_t, S_t)$, and transits into the next state $S_{t+1} \overset{d}{\sim} \mathcal{T}(\bullet \mid A_t, S_t)$ according to the transition function $\mathcal{T}$. Notice that both the reward and transition functions rely only on the current state-action pair (S_t, A_t), independent of the past data history. This ensures that the data satisfies the Markov assumption.

Policy and value. Let π denote a given target policy we wish to evaluate. We use $\mathbb{E}^\pi$ and $\mathbb{P}^\pi$ to denote the expectation and probability assuming the actions are chosen according to π at each time. The regular $\mathbb{E}$ and $\mathbb{P}$ without superscript are taking respect to the behavior policy b. Our objective lies in estimating the expected cumulative reward under π, denoted by $J(\pi) = \mathbb{E}^\pi \Big[\sum_{t=1}^{+\infty} \gamma^{t-1} R_t \Big]$ using the offline dataset generated under a different policy b. Additionally, denote V^π and Q^π as the state value function and state-action value function (better known as the Q-function), namely,

$$V^\pi(s) = \mathbb{E}^\pi \Big[\sum_{t=1}^{+\infty} \gamma^{t-1} R_t | S_1 = s \Big] \text{ and } Q^\pi(a, s) = \mathbb{E}^\pi \Big[\sum_{t=1}^{+\infty} \gamma^{t-1} R_t | S_1 = s, A_1 = a \Big]. \quad (1)$$

These functions are pivotal in developing value-based estimators, as described in Method 1 of Section 2.3. Moreover, we use π^* to denote the optimal policy that maximizes $J(\pi)$, i.e., $\pi^* \in \arg\max_\pi J(\pi)$, and write the optimal Q- and value functions Q^{π^*}, V^{π^*} as Q^*, V^* for brevity.

IS ratio. We also introduce the IS ratio $\rho^\pi(a, s) = \pi(a|s)/b(a|s)$, which quantifies the discrepancy between the target policy π and the behavior policy b. Furthermore, let $w^\pi(a, s)$ denote the MIS ratio $(1-\gamma) \sum_{t\geq 1} \gamma^{t-1} \mathbb{P}^\pi(S_t = s, A_t = a) / \lim_{t\to\infty} \mathbb{P}(S_t = s, A_t = a)$. Here, the numerator represents the discounted visitation probability under the target policy π, a crucial component in policy-based learning for estimating π^* (Sutton et al., 1999; Schulman et al., 2015). The denominator corresponds to the limiting state-action distribution under the behavior policy. These ratios are fundamental in constructing IS estimators, as detailed in Methods 2 and 3 of Section 2.3.

2.2 State abstractions for policy learning

Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \rho_0, \gamma \rangle$ be the ground MDP. A state abstraction ϕ is a mapping from the state space $\mathcal{S}$ to certain abstract state space $\mathcal{X} = \{\phi(s) : s \in \mathcal{S}\}$. Below, we review some commonly studied definitions of state abstraction designed for learning the optimal policy π^*; see Jiang (2018).

Definition 1 (**π^*-irrelevance**) *ϕ is π^*-irrelevant if there exists an optimal policy π^*, such that for any $s^{(1)}$, $s^{(2)} \in \mathcal{S}$ whenever $\phi(s^{(1)}) = \phi(s^{(2)})$, we have $\pi^*(a|s^{(1)}) = \pi^*(a|s^{(2)})$ for any $a \in \mathcal{A}$.*

Definition 2 (**Q^*-irrelevance**) *ϕ is Q^*-irrelevant if for any $s^{(1)}$, $s^{(2)} \in \mathcal{S}$ whenever $\phi(s^{(1)}) = \phi(s^{(2)})$, the optimal Q-function satisfies $Q^*(a, s^{(1)}) = Q^*(a, s^{(2)})$ for any $a \in \mathcal{A}$.*

Definitions 1 and 2 are easy to understand, requiring the optimal policy/Q-function to depend on a state s only through its abstraction $\phi(s)$. In practical terms, these definitions encourage the transformation of raw MDP data into a new sequence of state-action-reward triplets $(\phi(S), A, R)$ for policy learning. However, the transformed data may not necessarily satisfy the Markov assumption. This leads us to define the following model-irrelevance, which aims to preserve the MDP structure while ensuring π^*- and Q^*-irrelevance.

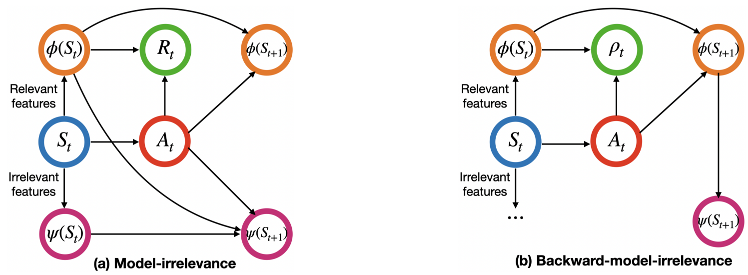


Figure 1: Illustrations of (a) model-irrelevance and (b) backward-model-irrelevance. ρ_t is a shorthand for $\rho^\pi(A_t, S_t)$ for any $t \geq 1$.

Definition 3 (Model-irrelevance) *ϕ is model-irrelevant if for any $s^{(1)}$, $s^{(2)} \in \mathcal{S}$ whenever $\phi(s^{(1)}) = \phi(s^{(2)})$, the following holds for any $a \in \mathcal{A}$, $s' \in \mathcal{S}$ and $x' \in \mathcal{X}$:*

$$\mathcal{R}(a, s^{(1)}) = \mathcal{R}(a, s^{(2)}) \ \ \textit{and} \sum_{s' \in \phi^{-1}(x')} \mathcal{T}(s'|a, s^{(1)}) = \sum_{s' \in \phi^{-1}(x')} \mathcal{T}(s'|a, s^{(2)}). \tag{2}$$

The first condition in (2) corresponds to "reward-irrelevance" whereas the second condition represents "transition-irrelevance". Consequently, Definition 3 defines a "model-based" abstraction, in contrast to "model-free" abstractions considered in Definitions 1 and 2. Notice that the term $\sum_{s' \in \phi^{-1}(x')} \mathcal{T}(s'|a, s)$ – appearing in the second equation of (2) – represents the probability of transitioning to $\phi(S') = x'$ in the abstract state space. Thus, the second condition essentially requires the abstract next state $\phi(S')$ to be conditionally independent of S given A and $\phi(S)$. Assuming S can be decomposed into the union of $\phi(S)$ and $\psi(S)$, which represent relevant features and irrelevant features, respectively. The condition implies that the evolution of those relevant features depends solely on themselves, independent of those irrelevant features. This ensures that the transformed data triplets $(\phi(S), A, R)$ remains an MDP. Meanwhile, the evolution of those irrelevant features may still depend on the relevant features; see Figure 1(a) for an illustration.

It is also known that model-irrelevance implies Q^*-irrelevance, which in turn implies π^*-irrelevance; see e.g., Theorem 2 in Li et al. (2006). Given that the transformed data remains an MDP under model-irrelevance, one can apply existing state-of-the-art RL algorithms to the abstract state space instead of the original ground space, leading to more effective learning of the optimal policy.

2.3 OPE methodologies

We focus on four OPE methods, covering the three families of estimators introduced in Section 1.1. Each method employs a specific formula to identify $J(\pi)$, which we detail below. The first method is a popular value-based approach – the Q-function-based method. The second and third methods are the two major IS estimators: SIS and MIS. The fourth method is a semi-parametrically efficient doubly robust method, double RL (DRL), known for achieving the smallest possible MSE among a broad class of OPE estimators (Kallus & Uehara, 2020, 2022).

Method 1 (Q-function-based method). For a given Q-function Q, define $f_1(Q)$ as the estimating function $\sum_{a \in \mathcal{A}} \pi(a|S_1) Q(a, S_1)$ with S_1 being the initial state. By (1) and the definition of $J(\pi)$, it is immediate to see that $J(\pi) = \mathbb{E}[f_1(Q^\pi)]$. This motivates the Q-function-based method which uses a plug-in estimator to approximate $\mathbb{E}[f_1(Q^\pi)]$ and thereby estimates $J(\pi)$. In particular, Q^π can be estimated by Q-learning type algorithms (e.g., fitted Q-evaluation, FQE, Le et al., 2019), and the expectation can be approximated based on the empirical initial state distribution.

Method 2 (Sequential importance sampling). For a given IS ratio ρ^π, let $\rho^\pi_{1:t}$ denote the cumulative IS ratio $\prod_{j=1}^t \rho^\pi(A_j, S_j)$. It follows from the change of measure theorem that the counterfactual reward $\mathbb{E}^\pi(R_t)$ is equivalent to $\mathbb{E}(\rho^\pi_{1:t} R_t)$ whose expectation is taken with respect to the offline data distribution. Assuming all trajectories in $\mathcal{D}$ terminate after a finite time T, this allows us to approximate $J(\pi)$ by $\mathbb{E}[f_2(\rho^\pi)]$ where $f_2(\rho^\pi) = \sum_{t=1}^T \gamma^{t-1} \rho^\pi_{1:t} R_t$. The approximation error is bounded by $O(\gamma^T)$, which decays exponentially fast with respect to T. SIS utilizes a plug-in estimator to initially estimate ρ^π (when the behavior policy is unknown), and subsequently employs this estimator, along with the empirical data distribution, to approximate $\mathbb{E}[f_2(\rho^\pi)]$. However, a notable limitation of this estimator is its rapidly increasing variance due to the use of the cumulative IS ratio $\rho^\pi_{1:t}$. Specifically, this variance tends to grow exponentially with respect to t, a phenomenon often referred to as *the curse of horizon* (Liu et al., 2018).

Method 3 (Marginalized importance sampling). The MIS estimator is designed to overcome the limitations of the SIS estimator. It breaks the curse of horizon by incorporating the structure of the MDP model. As noted previously, under the Markov assumption, the reward depends only on the current state-action pair, rather than the entire history. This insight allows us to replace the cumulative IS ratio with the MIS ratio, which depends solely on the current state-action pair. This modification considerably reduces variance because w^π is no longer history-dependent. Assuming the data trajectory is stationary over time – that is, all state-action-reward (S, A, R) triplets have the same distribution – it can be shown that $J(\pi) = \mathbb{E}[f_3(w^\pi)]$ where $f_3(w^\pi) = (1-\gamma)^{-1} w^\pi(A, S)R$ for any triplet (S, A, R). Both w^π and the expectation can be effectively estimated and approximated using offline data.

Method 4 (Double reinforcement learning). DRL combines Q-function-based method with MIS. Let $f_4(Q, w) = f_1(Q) + (1-\gamma)^{-1} w(A, S)[R + \gamma \sum_a \pi(a|S')Q(a, S') - Q(A, S)]$, where f_1 is defined in Method 1 and (S, A, R, S') denotes a state-action-reward-next-state tuple. Under the stationarity assumption, it can be shown that $J(\pi) = \mathbb{E}[f_4(Q, w)]$ when either $Q = Q^\pi$ or $w = w^\pi$ (Kallus & Uehara, 2022). DRL proposes to learn both Q^π and w^π from the data, employing these estimators to calculate $\mathbb{E}[f_4(Q, w)]$ and approximate the expectation with empirical data distribution. The resulting estimator benefits from double robustness: it is consistent when either Q^π or w^π is correctly specified.

3 Proposed state abstractions for policy evaluation

Here, we propose model-free (Section 3.1) and model-based irrelevance conditions (Section 3.2) for OPE, and analyze the OPE estimators under these conditions (Theorem 1, Theorem 2, Theorem 3). Motivated by this analysis, we propose our two-step procedure (Section 3.3).

3.1 Model-free irrelevance conditions

We first introduce several model-free irrelevance conditions tailored for OPE.

Definition 4 (π-irrelevance) *ϕ is π-irrelevant if for any $s^{(1)}, s^{(2)} \in \mathcal{S}$ whenever $\phi(s^{(1)}) = \phi(s^{(2)})$, we have $\pi(a|s^{(1)}) = \pi(a|s^{(2)})$ for any $a \in \mathcal{A}$.*

Definition 5 (Q^π-irrelevance) *ϕ is Q^π-irrelevant if for any $s^{(1)}, s^{(2)} \in \mathcal{S}$ whenever $\phi(s^{(1)}) = \phi(s^{(2)})$, we have $Q^\pi(a, s^{(1)}) = Q^\pi(a, s^{(2)})$ for any $a \in \mathcal{A}$.*

Definitions 4 and 5 are adaptations of Definitions 1 and 2 designed for policy evaluation, with the optimal policy π^* replaced by the target policy π. The following definitions are tailored for IS estimators (see Methods 2 and 3 in Section 2.3).

Definition 6 (ρ^π-irrelevance) *ϕ is ρ^π-irrelevant if for any $s^{(1)}, s^{(2)} \in \mathcal{S}$ whenever $\phi(s^{(1)}) = \phi(s^{(2)})$, we have $\rho^\pi(a, s^{(1)}) = \rho^\pi(a, s^{(2)})$ for any $a \in \mathcal{A}$.*

Definition 7 (w^π-irrelevance) *ϕ is w^π-irrelevant if for any $s^{(1)}, s^{(2)} \in \mathcal{S}$ whenever $\phi(s^{(1)}) = \phi(s^{(2)})$, we have $w^\pi(a, s^{(1)}) = w^\pi(a, s^{(2)})$ for any $a \in \mathcal{A}$.*

Based on the aforementioned definitions, we can immediately state the following theorem:

Theorem 1 (OPE under model-free irrelevance conditions) *Under Q^π-, ρ^π- or w^π-irrelevance, the corresponding methods remain valid when applied to the abstract state space:*

- *Under Q^π-irrelevance, the Q-function-based method (Method 1) remains valid, i.e., the Q-function Q^π_ϕ defined on the abstract state space satisfies $\mathbb{E}[f_1(Q^\pi)] = \mathbb{E}[f_1(Q^\pi_\phi)]$;*
- *Under ρ^π-irrelevance, SIS (Method 2) remains valid, i.e., the IS ratio ρ^π_ϕ defined on the abstract state space satisfies $\mathbb{E}[f_2(\rho^\pi)] = \mathbb{E}[f_2(\rho^\pi_\phi)]$;*

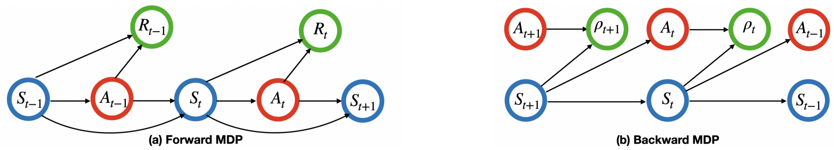


Figure 2: Illustrations of (a) the forward MDP model and (b) the backward MDP model.

- *Under w^π-irrelevance, MIS (Method 3) remains valid, i.e., the MIS ratio w^π_ϕ defined on the abstract state space satisfies $\mathbb{E}[f_3(w^\pi)] = \mathbb{E}[f_3(w^\pi_\phi)]$.*

Moreover, when ϕ satisfies either Q^π-irrelevance or w^π-irrelevance, DRL (Method 4) remains valid, i.e., Q^π_ϕ and w^π_ϕ defined on the abstract state space satisfy $\mathbb{E}[f_4(Q^\pi, w^\pi)] = \mathbb{E}[f_4(Q^\pi_\phi, w^\pi_\phi)]$.

Theorem 1 validates the four OPE methods presented in Section 2.3 when applied to the abstract state space, under the corresponding irrelevance conditions. Notably, DRL requires weaker irrelevance conditions compared to the Q-function-based method and MIS, owing to its inherent double robustness property. Nevertheless, methods for deriving abstractions that satisfy these conditions (particularly Q^π- and w^π-irrelevance) remain unclear. Furthermore, the state-action-reward triplets transformed via these abstractions $(\phi(S), A, R)$ might not maintain the MDP structure. This complicates the process of learning Q^π_ϕ and w^π_ϕ. These challenges motivate us to develop model-based irrelevance conditions in the subsequent section.

3.2 Model-based irrelevance conditions

To begin with, we discuss two perspectives of the data generated within the MDP framework; see Figure 2 for a graphical illustration.

1. The first perspective is the traditional **forward MDP** model with all state-action-reward triplets sequenced by time index. This yields the model-based irrelevance condition defined in Definition 3. We will discuss the relationship between this condition and Definitions 5-7 below.
2. The second perspective offers a backward view by reversing the time order. Specifically, due to the symmetric nature of the Markov assumption — implying that if the future is independent of the past given the present, the past must also be independent of the future given the present — the reversed state-action pairs also maintain the Markov property. Leveraging this property, we define another **backward MDP**, which forms the basis for deriving model-based conditions for achieving w^π-irrelevance and motivates the subsequent two-step procedure. This development represents one of our main contributions.

Forward MDP-based model-irrelevance. We first explore the relationship between the model-irrelevance given in Definition 3, and the notions of Q^π-, ρ^π- and w^π-irrelevance.

Theorem 2 (OPE under model-irrelevance) *Let ϕ denote a model-irrelevant abstraction.*

- *If ϕ is additionally π-irrelevant, then ϕ is also Q^π-irrelevant.*
- *While ϕ is not necessarily w^π-irrelevant, MIS (Method 3) remains valid when applied to the abstract state space. Indeed, the validity only requires reward-irrelevance (see the first part of* (2)*).*
- *While ϕ is not necessarily ρ^π-irrelevant, SIS (Method 2) remains valid when applied to the abstract state space if ϕ is additionally π-irrelevant.*
- *DRL (Method 4) remains valid when applied to the abstract state space.*

The first bullet point establishes the link between model-irrelevance and Q^π-irrelevance, thus proving the validity of the Q-function-based method when applied to the abstract state space. To satisfy Q^π-irrelevance, we need both model-irrelevance and π-irrelevance. In our implementation, we first adapt existing algorithms (Ha & Schmidhuber, 2018; François-Lavet et al., 2019; Gelada et al., 2019) to train a model-irrelevant abstraction ϕ, parameterized via deep neural networks. We next combine $\phi(s)$ with $\{\pi(a|s) : a \in \mathcal{A}\}$ to obtain a new abstraction $\phi_{for}(s)$. This augmentation ensures $\phi_{for}(s)$ is π-irrelevant, and hence Q^π-irrelevant. Refer to Appendix B.1 for the detailed procedures.

The last three bullet points prove the validity of the SIS, MIS and DRL, despite ϕ being neither w^π-irrelevant nor ρ^π-irrelevant. By definition, ρ^π-irrelevance can be achieved by selecting state features that adequately predict the IS ratio. However, methods for constructing w^π-irrelevant abstractions remain less clear. In the following, we introduce a backward MDP model-based irrelevance condition that ensures w^π-irrelevance. We also note that findings similar to those in the first two bullet points have previously been documented in Li et al. (2006) and Pavse & Hanna (2023), respectively. However, the properties of SIS and DRL estimators under model-irrelevance conditions as summarized in our last two bullet points, remain unexplored in the existing literature.

Backward MDP-based model-irrelevance. To illustrate the rationale behind the proposed model-based abstraction, we introduce the backward MDP model by reversing the time index. Under the (forward) MDP model assumption described in Section 2.1 and that the behavior policy b is not history-dependent, actions and states following S_t are independent of those occurred prior to the realization of S_t. Accordingly, (S_{t-1}, A_{t-1}) is conditionally independent of $\{(S_k, A_k)\}_{k>t}$ given S_t. Recall that T corresponds to the termination time of trajectories in $\mathcal{D}$. We define a time-reversed process consisting of state-action-reward triplets $\{(S_t, A_t, \rho^\pi(A_t, S_t)) : t = T, \ldots, 1\}$. Its dynamics is described as follows (see also Figure 2(b) for the configuration):

- **State-action transition**: Due to the aforementioned Markov property, the transition of the past state S_{t+1} in the reversed process (future state in the original process) into the current state S_t is independent of the past action A_{t+1} in the reversed process (future action in the original process) while the behavior policy that generates A_t depends on both the current state S_t and the past state S_{t+1} in the reversed process. This yields the time-reversed state-action transition function $\mathbb{P}(A_t = a, S_t = s | S_{t+1})$.
- **Reward generation**: For each state-action pair (S_t, A_t), we manually set the reward to the IS ratio $\rho^\pi(A_t, S_t)$, which plays a crucial role in constructing IS estimators.

Given this MDP, analogous to Definition 3, our objective is to identify a state abstraction that is crucial for predicting the reward (e.g., the IS ratio) and the reversed transition function. We provide the formal definition of the proposed backward MDP-based model-irrelevance (short for backward-model-irrelevance) below.

Definition 8 (Backward-model-irrelevance) *ϕ is backward-model-irrelevant if for any $s^{(1)}, s^{(2)} \in \mathcal{S}$ whenever $\phi(s^{(1)}) = \phi(s^{(2)})$, the followings hold for any $a \in \mathcal{A}$, $x \in \mathcal{X}$ and $t \in \mathbb{N}^+$:*

$$(i) \rho^\pi(a, s^{(1)}) = \rho^\pi(a, s^{(2)});$$

$$(ii) \sum_{s \in \phi^{-1}(x)} \mathbb{P}(A_t = a, S_t = s | S_{t+1} = s^{(1)}) = \sum_{s \in \phi^{-1}(x)} \mathbb{P}(A_t = a, S_t = s | S_{t+1} = s^{(2)}). \quad (3)$$

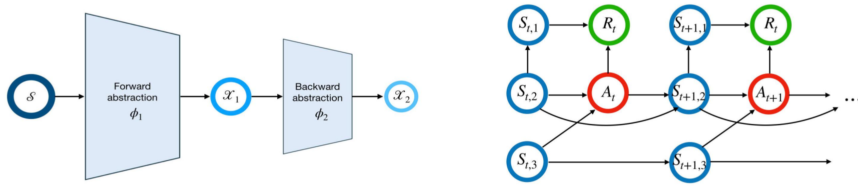
The conditions of backward-model-irrelevance are similar to those specified for model-irrelevance outlined in Definition 3. The first condition (i) essentially requires reward-irrelevance, i.e., ρ^π-irrelevance, in the backward MDP. The second condition in equation (3) is equivalent to the conditional independence assumption between the pair $(A_t, \phi(S_t))$ and S_{t+1} given $\phi(S_{t+1})$. As previously assumed, S_t can be decomposed into the union of relevant features $\phi(S_t)$ and irrelevant features $\psi(S_t)$, leading to the following factorization:

$$\mathbb{P}(S_{t+1} = s' | A_t, \phi(S_t)) = \mathbb{P}(\psi(S_{t+1}) = \psi(s') | \phi(S_{t+1}) = \phi(s')) \mathbb{P}(\phi(S_{t+1}) = \phi(s') | A_t, \phi(S_t)).$$

This indicates a two-step transition in the forward model: initially from $(\phi(S_t), A_t)$ to $\phi(S_{t+1})$, and then from $\phi(S_{t+1})$ to $\psi(S_{t+1})$. Importantly, the generation of $\psi(S_{t+1})$ in the second step is conditionally independent of A_t and $\phi(S_t)$. Consequently, ϕ extracts state representations that are influenced either by past actions or past relevant features; see Figure 1(b) for an illustration. Combined with ρ^π-irrelevance, this ensures that all information contained within the historical IS ratios $\{\rho^\pi(A_k, S_k)\}_{k<t}$ can be effectively summarized using a single A_{t-1} and the abstract state $\phi(S_{t-1})$, thus achieving w^π-irrelevance (see Theorem 3 below).

Theorem 3 (OPE under backward-model-irrelevance) *Assume ϕ is backward-model-irrelevant.*

- *ϕ is both ρ^π-irrelevant and w^π-irrelevant.*
- *While ϕ is not necessarily Q^π-irrelevant, the Q-function-based method (Method 1) remains valid when applied to the abstract state space.*



(a) Two-step procedure

(b) An MDP example

Figure 3: Illustrations of (a) the two-step procedure and (b) an MDP with three groups of state variables, denoted by $\{S_{t,1}\}_t$, $\{S_{t,2}\}_t$ and $\{S_{t,3}\}_t$.

- *DRL (Method 4) remains valid when applied to the abstract state space.*

The first bullet point in Theorem 3 validates the two IS methods when applied to the abstract state space under the proposed backward-model-irrelevance, whereas the last two bullet points validate the Q-function-based method and DRL.

To conclude this section, we draw a connection between the proposed backward-model-irrelevant abstraction for OPE and the Markov state abstraction (MSA) developed by Allen et al. (2021) for policy learning. MSA impose two conditions: (i) inverse-model-irrelevance, which requires A_t to be conditionally independent of S_t and S_{t+1} given $\phi(S_t)$ and $\phi(S_{t+1})$; (ii) density-ratio-irrelevance, which requires $\phi(S_t)$ to be conditionally independent of S_{t+1} given $\phi(S_{t+1})$. For effective policy learning, MSA requires both conditions to hold in data generating processes following a diverse range of behavior policies. When restricting them to one behavior policy, the two conditions are closely related to our backward-model-irrelevance. In particular, they imply our proposed condition in (3) whereas (3) in turn yields density-ratio-irrelevance. This allows us to adapt their algorithm to train state abstractions that satisfy backward-model-irrelevance; see Appendix B.2 for details.

3.3 Two-step procedure for forward and backward state abstraction

The proposed two-step procedure proceeds as follows (see Figure 3(a) for a visualization):

1. **Forward abstraction**: learn an abstraction ϕ_1 from the ground state space $\mathcal{S} = \mathcal{X}_0$ to $\mathcal{X}_1$ using the data triplets (S, A, R) that is both (forward)-model-irrelevant and π-irrelevant.
2. **Backward abstraction**: Learn an abstraction ϕ_2 from the abstract state space $\mathcal{X}_1$ to $\mathcal{X}_2$ using the data triplets $(\phi_1(S), A, R)$ that is backward-model-irrelevant.
3. **Output** $\mathcal{X}_2$ for off-policy evaluation.

To summarize, our approach sequentially applies the forward and backward abstraction on the state obtained from the previous iteration, progressively reducing state cardinality. To elaborate the usefulness of the two-step procedure in reducing state cardinality, we first analyze a toy example.

A toy example: Consider an MDP where the state variables can be classified into three groups, depicted in Figure 3(b). For this example, we focus on a specific type of state abstraction known as variable selection, which selects a sub-vector from the original state. Key observations from this example are as follows: (i) The reward depends on the state only through the first group of variables; (ii) The evolution of the first group of variables depends only on the second group, and this dependency is indirect. Specifically, the second group evolves first at each time step and subsequently influences the first group; (iii) The second and third groups in the MDP evolve independently, each relying solely on their own previous states; (iv) The behavior policy depends only on the last two groups; (v) Only the second group of variables is directly influenced by the previous action.

According to (i), selecting the first group of variables achieves reward-irrelevance. Combined with (ii) and (iii), choosing the first two groups achieves model-irrelevance. Assuming the target policy is agnostic to the state, the proposed forward abstraction will select the first two groups of variables.

According to (iv) and that the target policy is state-agnostic, selecting the last two groups attains ρ^{π}-irrelevance. Meanwhile, according to (ii) and (v), selecting these variables also achieves backward-model-irrelevance. Thus, the proposed backward abstraction will select the last two groups.

In the two-step procedure, the forward abstraction first eliminates the third group of variables. Given conditions (ii)-(v), selecting just the second group suffices to achieve backward-model-irrelevance, leading to the elimination of the first group in the subsequent backward abstraction. After two iterations, the procedure produces only one group of variables, demonstrating its efficiency in reducing dimensions compared to using either forward or backward abstraction alone.

In more complex scenarios, each abstraction guarantees that the cardinality of the state space does not increase, effectively maintaining or reducing complexity. The reduction is more likely because forward and backward abstractions, as illustrated in Figures 1(a) and (b), differ by definition. Meanwhile, according to Theorems 2 and 3, the post-abstraction-OPE remains valid for any of the four methods.

Theorem 4 (The two-step procedure) *The four OPE methods remain valid when applied to the abstracted state produced by the proposed two-step procedure.*

Finally, we note that one may further consider an iterative procedure that alternates between forward and backward abstractions. However, it remains unclear whether these methods have guarantees.

4 Numerical experiments

Method. We investigate the finite sample performance of our proposed methods (details in Appendix B), the forward, backward and two-step procedures.

Comparisons. We compare the proposed abstraction obtained via the two-step procedure (denoted by ‘two-step’), single-iteration forward (‘forward’) and backward (‘backward’) abstractions against Markov state abstraction (Allen et al., 2021) (‘Markov’) and a reconstruction-based abstraction (Lange & Riedmiller, 2010) (‘auto-encoder’). Each abstraction’s performance is tested using FQE (Le et al., 2019) applied to the abstract state space. We also report the performance of a baseline FQE applied to the unabstracted, ground state space (‘FQE’).

Environments. We consider two environments from OpenAI Gym (Brockman et al., 2016), “CartPole-v0” and “LunarLander-v2”, with original state dimensions of 4 and 8, respectively. For each environment, we manually include 296 and 292 irrelevant variables in the state, leading to a challenging 300-dimensional system. Refer to Appendix C for more details about these environments.

Results. We report the MSEs and biases of different post-abstraction-OPE estimators and those of the baseline FQE estimator without abstraction in Figure 4 and Figure C.1 in Appendix C. We summarize our findings as follows. First, the proposed two-step method outperforms other baseline methods, with the smallest MSE and absolute bias in all cases. Since ‘Markov’ and ‘auto-encoder’ are types of model-irrelevant abstractions, these comparisons demonstrate the advantages of the proposed two-step method over single-iteration forward and backward procedures. Second, both figures indicate that the baseline FQE applied to the ground state space performs the worst among all cases. This demonstrates the usefulness of state abstractions for OPE.

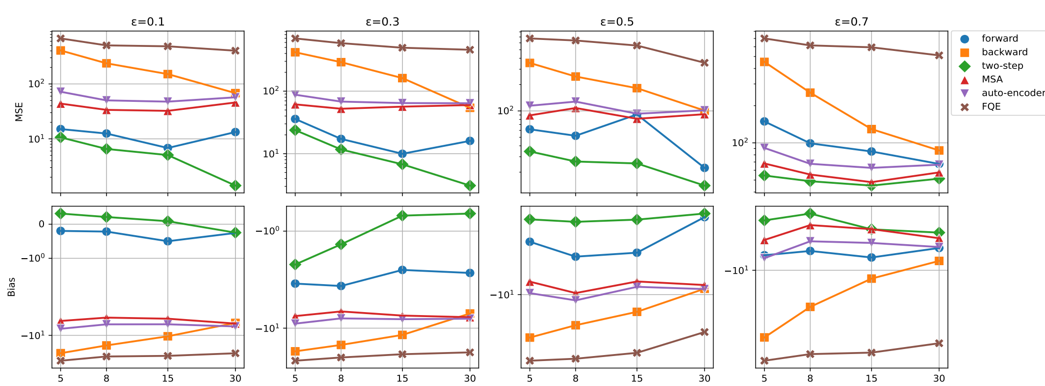


Figure 4: MSEs and biases of FQE estimators when applied to ground and abstract state spaces with various abstractions. The behavior policy is ϵ-greedy with respect to the target policy, with $\epsilon = 0.1, 0.3, 0.5, 0.7$ from left to right.

References

Abel, D. A theory of abstraction in reinforcement learning. *arXiv preprint arXiv:2203.00397*, 2022.

Abel, D., Hershkowitz, D., and Littman, M. Near optimal behavior via approximate state abstraction. In *International Conference on Machine Learning*, pp. 2915–2923, 2016.

Allen, C., Parikh, N., Gottesman, O., and Konidaris, G. Learning Markov state abstractions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 8229–8241, 2021.

Austin, P. C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011.

Belloni, A., Chernozhukov, V., and Hansen, C. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Castro, P. S. Scalable methods for computing state similarity in deterministic Markov decision processes. In *AAAI Conference on Artificial Intelligence*, pp. 10069–10076, 2020.

Chapelle, O. and Li, L. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, pp. 2249–2257, 2011.

Chen, X. and Qi, Z. On well-posedness and minimax optimal rates of nonparametric Q-function estimation in off-policy evaluation. In *International Conference on Machine Learning*, pp. 3558–3582, 2022.

Dai, B., Nachum, O., Chow, Y., Li, L., Szepesvári, C., and Schuurmans, D. CoinDICE: Off-policy confidence interval estimation. In *Advances in Neural Information Processing Systems*, pp. 9398–9411, 2020.

De Luna, X., Waernbaum, I., and Richardson, T. S. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875, 2011.

Dean, T. and Givan, R. Model minimization in Markov decision processes. In *Conference on Artificial Intelligence / Conference on Innovative Applications of Artificial Intelligence*, pp. 106–111, 1997.

Dudík, M., Erhan, D., Langford, J., and Li, L. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1447–1456, 2018.

Feng, Y., Ren, T., Tang, Z., and Liu, Q. Accountable off-policy evaluation with kernel Bellman statistics. In *International Conference on Machine Learning*, pp. 3102–3111, 2020.

Ferns, N., Panangaden, P., and Precup, D. Metrics for finite Markov decision processes. In *Conference on Uncertainty in Artificial Intelligence*, pp. 162–169, 2004.

Ferns, N., Panangaden, P., and Precup, D. Bisimulation metrics for continuous Markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011.

François-Lavet, V., Bengio, Y., Precup, D., and Pineau, J. Combined reinforcement learning via abstract representations. In *AAAI Conference on Artificial Intelligence*, pp. 3582–3589, 2019.

Fu, J., Norouzi, M., Nachum, O., Tucker, G., Novikov, A., Yang, M., Zhang, M. R., Chen, Y., Kumar, A., Paduraru, C., et al. Benchmarks for deep off-policy evaluation. In *International Conference on Learning Representations*, 2020.

Gelada, C., Kumar, S., Buckman, J., Nachum, O., and Bellemare, M. G. DeepMDP: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, pp. 2170–2179, 2019.

Givan, R., Dean, T., and Greig, M. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1-2):163–223, 2003.

Glymour, M. M., Weuve, J., and Chen, J. T. Methodological challenges in causal research on racial and ethnic patterns of cognitive trajectories: measurement, selection, and bias. *Neuropsychology Review*, 18:194–213, 2008.

Greenland, S., Pearl, J., and Robins, J. M. Confounding and collapsibility in causal inference. *Statistical science*, 14(1):29–46, 1999.

Guo, F. R. and Zhao, Q. Confounder selection via iterative graph expansion. *arXiv preprint arXiv:2309.06053*, 2023.

Guo, F. R., Lundborg, A. R., and Zhao, Q. Confounder selection: Objectives and approaches. *arXiv preprint arXiv:2208.13871*, 2022.

Ha, D. and Schmidhuber, J. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, pp. 2455–2467, 2018.

Hanna, J., Niekum, S., and Stone, P. Importance sampling policy evaluation with an estimated behavior policy. In *International Conference on Machine Learning*, pp. 2605–2613, 2019.

Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvari, C., and Wang, M. Bootstrapping fitted Q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pp. 4074–4084, 2021.

Hernán, M. A. and Robins, J. M. Causal inference, 2010.

Hernán, M. A. and Robins, J. M. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764, 2016.

Hu, Y. and Wager, S. Off-policy evaluation in partially observed Markov decision processes under sequential ignorability. *The Annals of Statistics*, 51(4):1561–1585, 2023.

Jiang, H., Dai, B., Yang, M., Zhao, T., and Wei, W. Towards automatic evaluation of dialog systems: A model-free off-policy evaluation approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7419–7451, 2021.

Jiang, N. Notes on state abstractions, 2018.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661, 2016.

Jong, N. K. and Stone, P. State abstraction discovery from irrelevant state variables. In *International Joint Conference on Artificial Intelligence*, pp. 752–757, 2005.

Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 70(6):3282–3302, 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Koch, B., Vock, D. M., Wolfson, J., and Vock, L. B. Variable selection and estimation in causal inference using Bayesian spike and slab priors. *Statistical Methods in Medical Research*, 29(9): 2445–2469, 2020.

Lange, S. and Riedmiller, M. Deep auto-encoder neural networks in reinforcement learning. In *International Joint Conference on Neural Networks*, pp. 1–8, 2010.

Laskin, M., Srinivas, A., and Abbeel, P. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pp. 5639–5650, 2020.

Le, H., Voloshin, C., and Yue, Y. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712, 2019.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Li, G., Wu, W., Chi, Y., Ma, C., Rinaldo, A., and Wei, Y. High-probability sample complexities for policy evaluation with linear function approximation. *IEEE Transactions on Information Theory*, 2024.

Li, L., Walsh, T. J., and Littman, M. L. Towards a unified theory of state abstraction for MDPs. *AI&M*, 1(2):3, 2006.

Liao, P., Klasnja, P., and Murphy, S. Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, 116(533):382–391, 2021.

Liao, P., Qi, Z., Wan, R., Klasnja, P., and Murphy, S. A. Batch policy learning in average reward Markov decision processes. *Annals of Statistics*, 50(6):3364, 2022.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5361–5371, 2018.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. Estimating dynamic treatment regimes in mobile health using V-learning. *Journal of the American Statistical Association*, 115:692–706, 2019.

Ma, T., Cai, H., Qi, Z., Shi, C., and Laber, E. B. Sequential knockoffs for variable selection in reinforcement learning. *arXiv preprint arXiv:2303.14281*, 2023.

Mandel, T., Liu, Y.-E., Levine, S., Brunskill, E., and Popovic, Z. Offline policy evaluation across representations with applications to educational games. In *International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1077–1084, 2014.

Murphy, S. A., van der Laan, M. J., Robins, J. M., and Group, C. P. P. R. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Nachum, O., Chow, Y., Dai, B., and Li, L. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing systems*, pp. 2318–2328, 2019.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pp. 2778–2787, 2017.

Pavse, B. S. and Hanna, J. P. Scaling marginalized importance sampling to high-dimensional state-spaces via state abstraction. In *AAAI Conference on Artificial Intelligence*, pp. 9417–9425, 2023.

Pearl, J. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009. ISBN 978-0-521-89560-6. doi: 10.1017/CBO9780511803161.

Persson, E., Häggström, J., Waernbaum, I., and de Luna, X. Data-driven algorithms for dimension reduction in causal inference. *Computational statistics & data analysis*, 105:280–292, 2017.

Precup, D. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning*, pp. 759–766, 2000.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Ravindran, B. *An algebraic approach to abstraction in reinforcement learning*. University of Massachusetts Amherst, 2004.

Robins, J. M. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pp. 69–117. Springer, 1997.

Rubin, D. B. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, 28(9):1420–1423, 2009.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897, 2015.

Shelhamer, E., Mahmoudieh, P., Argus, M., and Darrell, T. Loss is its own reward: Self-supervision for reinforcement learning. *arXiv preprint arXiv:1612.07307*, 2016.

Shi, C., Wan, R., Chernozhukov, V., and Song, R. Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning*, pp. 9580–9591, 2021.

Shi, C., Zhang, S., Lu, W., and Song, R. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society Series B*, 84(3): 765–793, 2022.

Shortreed, S. M. and Ertefaie, A. Outcome-adaptive Lasso: variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

Singh, S., Jaakkola, T., and Jordan, M. Reinforcement learning with soft state aggregation. In *Advances in Neural Information Processing Systems*, pp. 361–368, 1994.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pp. 1057–1063, 1999.

Sutton, R. S., Szepesvári, C., and Maei, H. R. A convergent $O(n)$ algorithm for off-policy temporal-difference learning with linear function approximation. In *Advances in Neural Information Processing Systems*, pp. 1609–1616, 2008.

Tang, Z., Feng, Y., Li, L., Zhou, D., and Liu, Q. Doubly robust bias reduction in infinite horizon off-policy estimation. In *International Conference on Learning Representations*, 2020.

Tangkaratt, V., Morimoto, J., and Sugiyama, M. Model-based reinforcement learning with dimension reduction. *Neural Networks*, 84:1–16, 2016.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148, 2016.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. High-confidence off-policy evaluation. In *AAAI Conference on Artificial Intelligence*, pp. 3000–3006, 2015.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668, 2020.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline RL in low-rank MDPs. In *International Conference on Learning Representations*, 2021.

Uehara, M., Shi, C., and Kallus, N. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

Vander Weele, T. J. and Shpitser, I. A new criterion for confounder selection. *Biometrics*, 67(4): 1406–1413, 2011.

VanderWeele, T. J. Principles of confounder selection. *European Journal of Epidemiology*, 34: 211–219, 2019.

Voloshin, C., Le, H. M., Jiang, N., and Yue, Y. Empirical study of off-policy policy evaluation for reinforcement learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Wang, J., Qi, Z., and Wong, R. K. Projected state-action balancing weights for offline reinforcement learning. *The Annals of Statistics*, 51(4):1639–1665, 2023.

Wang, L., Laber, E. B., and Witkiewitz, K. Sufficient Markov decision processes with alternating deep neural networks. *arXiv preprint arXiv:1704.07531*, 2017.

Xie, C., Yang, W., and Zhang, Z. Semiparametrically efficient off-policy evaluation in linear Markov decision processes. In *International Conference on Machine Learning*, pp. 38227–38257, 2023.

Xie, T., Ma, Y., and Wang, Y.-X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, pp. 9668–9678, 2019.

Yin, M. and Wang, Y.-X. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3948–3958, 2020.

Zhang, A., McAllister, R. T., Calandra, R., Gal, Y., and Levine, S. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2020.

Zhang, B. and Zhang, M. Variable selection for estimating the optimal treatment regimes in the presence of a large number of covariates. *The Annals of Applied Statistics*, 12(4):2335–2358, 2018.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694, 2013.

Appendix

This appendix is structured as follows: Section A introduces additional related works on confounder selection in causal inference. The implementation details of the proposed state abstraction are discussed in Section B. Additional information concerning the environments and computing resources utilized is presented in Section C. The limitations of our method are discussed in Section D. All technical proofs can be found in Section E.

A Confounder selection in causal inference

Broadly speaking, confounding refers to the problem that even if two variables are not causes of each other, they may exhibit statistical association due to common causes. Controlling for confounding is a central problem in the design of observational studies, and many criteria for confounder selection have been proposed in the literature. A commonly adopted criterion is the "common cause heuristic", where the user only controls for covariates that are related to both the treatment and the outcome (Glymour et al., 2008; Austin, 2011; Shortreed & Ertefaie, 2017; Koch et al., 2020). Another widely used criterion is to simply use all covariates that are observed before the treatment in time (Rubin, 2009; Hernán & Robins, 2010, 2016). However, both of these approaches are not guaranteed to find a set of covariates that are sufficient to control for confounding. From a graphical perspective, confounder selection is essentially about finding a set of covariates that block all "back-door" paths (Pearl, 2009), but this requires full structural knowledge about the causal relationship between the variables which is often not possible. This motivated some methods that only require partial structural knowledge (Vander Weele & Shpitser, 2011; VanderWeele, 2019; Guo & Zhao, 2023). All the aforementioned methods need substantive knowledge about the treatment, outcome, and covariates. Other methods use statistical tests (usually of conditional independence) to trim a set of covariates that are assumed to control for confounding (Robins, 1997; Greenland et al., 1999; Hernán & Robins, 2010; De Luna et al., 2011; Belloni et al., 2014; Persson et al., 2017). The reader is referred to Guo et al. (2022) for a recent survey of objectives and approaches for confounder selection.

Confounder selection can be considered as a special example of our problem under certain conditions: (i) The state transition is independent, effectively transforming the MDP into a contextual bandit; (ii) The action space is binary, with the target policy consistently assigning either action 0 or action 1, aimed at assessing the average treatment effect; (iii) State abstractions are confined to variable selections. While our proposed two-step procedure shares similar spirits with the aforementioned algorithms, it addresses a more complex problem involving state transitions. Additionally, our focus is on abstraction that facilitates the engineering of new feature vectors, rather than merely selecting a subset of existing ones.

B Implementation details

In this section, we present implementation details for forward abstraction (Section B.1) and backward abstraction (Section B.2).

B.1 Implementation details for forward abstraction

We provide details for implementing the proposed forward abstraction in this subsection. We use deep neural networks to parameterize the forward abstraction and estimate the parameters by minimzing the following loss function:

$$\alpha_1\mathcal{L}_r + \beta_1\mathcal{L}_{\mathcal{T}} + \delta_1\mathcal{L}_Q + \lambda_1\mathcal{L}_{penalty}, \tag{B.1}$$

where $\mathcal{L}_r$, $\mathcal{L}_{\mathcal{T}}$ and $\mathcal{L}_Q$ are the loss functions detailed below, $\mathcal{L}_{penalty}$ is a penalty term, and $\alpha_1, \beta_1, \delta_1, \lambda_1$ are positive constant hyper-parameters whose values are reported in Table B.1.

By definition, the forward abstraction is required to achieve both model-irrelevance and π-irrelevance. As discussed in Section 3.2, our approach is to learn a model-irrelevant abstraction, denoted as ϕ, and then concatenate it with $\{\pi(a|\bullet) : a \in \mathcal{A}\}$. We denote the concatenated abstraction by ϕ_{for}.

We next detail the loss functions and the penalty term. The first two losses $\mathcal{L}_r$ and $\mathcal{L}_\mathcal{T}$ are to ensure reward-irrelevance and transition-irrelevance, respectively,

$$\mathcal{L}_r = \frac{1}{|\mathcal{D}|} \sum_{(S,A,R)\in\mathcal{D}} \big[R - \mathcal{R}_\phi\big(A, \phi(S)\big)\big]^2, \; \mathcal{L}_\mathcal{T} = \frac{1}{|\mathcal{D}|} \sum_{(S,A,S')\in\mathcal{D}} \|\mathcal{T}_\phi\big(A, \phi(S)\big) - \phi(S')\|_2^2,$$

where $\mathcal{R}_{\phi_0}$ and $\mathcal{T}_{\phi_0}$ are the estimated reward and transition functions applied to the abstract state space parameterized by deep neural networks as well, and $|\mathcal{D}|$ is the cardinality of the dataset $\mathcal{D}$.

The inclusion of the third loss function, $\mathcal{L}_Q$, is motivated by the demonstrated benefits of utilizing model-free objectives to guide the training of state abstractions in policy learning, as evidenced by Gelada et al. (2019); Ha & Schmidhuber (2018); François-Lavet et al. (2019). Given our interest in OPE, we integrate the following FQE loss into the objective function,

$$\mathcal{L}_Q = \frac{1}{|\mathcal{D}|} \sum_{(S,A,R,S')\in\mathcal{D}} \Big[R + \gamma \sum_{a\in\mathcal{A}} \pi(a|S') Q^-\big(\phi_{for}(S'), a\big) - Q\big(\phi_{for}(S), A\big)\Big]^2,$$

where Q^- and Q represent the estimated $Q^\pi_{\phi_{for}}$ function applied to the abstract state space during the previous and current iterations, respectively.

The above objectives allow us to effectively train forward abstractions. However, a potential concern is that the resulting abstraction and transition can collapse to some constant x_0 such that $\phi_{for}(S) \to x_0, \;\; \forall S \in \mathcal{S}$. To address this limitation, we include the following penalty function of two randomly drawn states to promote diversity in the abstractions:

$$\mathcal{L}_c = \frac{1}{|\mathcal{D}|(|\mathcal{D}|-1)} \sum_{S,\tilde{S}\in\mathcal{D}, S\neq\tilde{S}} \exp(-C_0\|\widehat{\phi}(S) - \widehat{\phi}(\tilde{S})\|_2)$$

for some positive scaling constant C_0, and $\widehat{\phi}(s)$ is the estimated abstract state from transition function. $\widehat{\phi}(\tilde{s})$ can be achieved by shuffling $\widehat{\phi}(s')$ from pairs (s, s') in the batch. Additionally, we add another penalty to penalize consecutive abstract states for being more than some predefined distance d_0 away from each other,

$$\mathcal{L}_s = \frac{1}{|\mathcal{D}|} \sum_{(S,S')\in\mathcal{D}} C_1[\|\phi_{for}(S) - \phi_{for}(S')\|_2 - d_0]^2,$$

for some positive constant C_1. These components combine into the final penalty function:

$$\mathcal{L}_{penalty} = \mathcal{L}_s + \mathcal{L}_c.$$

The forward model architecture is as follow:

```
    Forward_model(
  (encoder): Encoder_linear(
    (activation): ReLU()
    (encoder_net): Sequential(
      (0): Linear(in_features=300, out_features=64, bias=True)
      (1): ReLU()
      (2): Linear(in_features=64, out_features=64, bias=True)
      (3): ReLU()
      (4): Dropout(p=0.2, inplace=False)
      (5): Linear(in_features=64, out_features=64, bias=True)
      (6): ReLU()
      (7): Dropout(p=0.2, inplace=False)
      (8): Linear(in_features=64, out_features=100, bias=True)
    )
  )
  (transition): Transition(
    (activation): ReLU()
    (T_net): Sequential(
      (0): Linear(in_features=100, out_features=64, bias=True)
```

```
      (1): ReLU()
      (2): Linear(in_features=64, out_features=64, bias=True)
      (3): ReLU()
      (4): Dropout(p=0.2, inplace=False)
      (5): Linear(in_features=64, out_features=64, bias=True)
    )
    (lstm): LSTMCell(64, 128)
    (tanh): Tanh()
  )
  (reward): Reward(
    (activation): ReLU()
    (reward_net): Sequential(
      (0): Linear(in_features=100, out_features=64, bias=True)
      (1): ReLU()
      (2): Linear(in_features=64, out_features=64, bias=True)
      (3): ReLU()
      (4): Dropout(p=0.2, inplace=False)
      (5): Linear(in_features=64, out_features=64, bias=True)
      (6): ReLU()
      (7): Dropout(p=0.2, inplace=False)
      (8): Linear(in_features=64, out_features=64, bias=True)
      (9): ReLU()
      (10): Dropout(p=0.2, inplace=False)
      (11): Linear(in_features=64, out_features=64, bias=True)
      (12): ReLU()
      (13): Dropout(p=0.2, inplace=False)
      (14): Linear(in_features=64, out_features=2, bias=True)
    )
  )
  (FQE): FQE(
    (activation): ReLU()
    (action_net): Sequential(
      (0): Linear(in_features=1, out_features=16, bias=True)
      (1): ReLU()
      (2): Linear(in_features=16, out_features=100, bias=True)
    )
    (xa_net): Linear(in_features=200, out_features=100, bias=True)
    (FQE_net): Sequential(
      (0): Linear(in_features=100, out_features=64, bias=True)
      (1): ReLU()
      (2): Linear(in_features=64, out_features=64, bias=True)
      (3): ReLU()
      (4): Dropout(p=0.2, inplace=False)
      (5): Linear(in_features=64, out_features=64, bias=True)
      (6): ReLU()
      (7): Dropout(p=0.2, inplace=False)
      (8): Linear(in_features=64, out_features=2, bias=True)
    )
  )
)
```

B.2 Implementation details for backward abstraction

We provide details for implementing the proposed backward abstraction in this subsection. Similar to Section B.1, we use deep neural networks to parameterize the abstraction ϕ_{back} and estimate the parameters by solving the following loss function,

$$\alpha_2\mathcal{L}_\rho + \beta_2\mathcal{L}_{ratio} + \delta_2\mathcal{L}_{inv} + \lambda_2\mathcal{L}_s,$$

where $\alpha_2, \beta_2, \delta_2, \lambda_2$ are positive hyper-parameters specified in Table B.1.

Table B.1: Hyper-parameters information. m is the input feature dimension, and $**$ means no value.

Environment	Hyper-parameters	Values	Hyper-parameters	Values
CartPole-v0	α_1	1	α_2	1
	β_1	1	β_2	1
	γ_1	1	γ_2	1
	λ_1	$\min(1, \frac{20}{m})$	λ_2	$\min(1, \frac{10}{m})$
	C_0	1	C_0	$**$
	C_1	1	C_1	1
	d_0	$0.15m$	d_0	$0.15m$
LunarLander-v2	α_1	1	α_2	1
	β_1	1	β_2	1
	γ_1	1	γ_2	1
	λ_1	$\min(1, \frac{20}{m})$	λ_2	$\min(1, \frac{20}{m})$
	C_0	1	C_0	$**$
	C_1	1	C_1	1
	d_0	$0.15m$	d_0	$0.15m$

Recall that backward-model-irrelevance requires both ρ^π-irrelevance (Definition 6) and (3). The first loss function $\mathcal{L}_\rho$ is designed to enforce ρ^π-irrelevance, specified as

$$\mathcal{L}_\rho = \frac{1}{|\mathcal{D}|} \sum_{(S,A)\in\mathcal{D}} \left[\widehat{\rho}^\pi(A,S) - \rho^\pi_{\phi_{back}}\big(A, \phi_{back}(S)\big)\right]^2,$$

where $\widehat{\rho}^\pi$ denotes some consistent estimator of the IS ratio. Note that in two-step procedure, we should replace $\widehat{\rho}^\pi(A,S)$ by:

$$\widehat{\rho}^\pi_{for}(A, \phi_{for}(S)) = \frac{\pi_{\phi_{for}}(A|\phi_{for}(S))}{\widehat{b}(A|\phi_{for}(S))} = \frac{\pi(A|S)}{\widehat{b}(A|\phi_{for}(S))},$$

where $\widehat{b}$ is estimated from the abstracted experiences and $\pi(A|S)$ keeps static due to the π-irrelevance property of forward abstraction.

As commented in Section 3.2, the second condition of (3) holds by satisfying the conditional independence assumption between $(A_t, \phi(S_t))$ and S_{t+1} given $\phi(S_{t+1})$. By Bayesian formula, we can show that it is satisfied by the inverse-model-irrelevance and density-ratio-irrelevance when setting the learning policy π to b. This motivates us to leverage the two objectives $\mathcal{L}_{inv}$ and $\mathcal{L}_{ratio}$ used by Allen et al. (2021) for training MSA. More details regarding these losses can be found in Section 5 of Allen et al. (2021). Note that to obtain non-sequential states $(s, \tilde{s})$ used in L_{ratio}, we flip s' in the pairs (s, s') in each batch instead of shuffling.

Finally, $\mathcal{L}_s$ corresponds to the smoothness penalty introduced in Section B.1. The backward model architecture is:

```
    Backward_model(
  (encoder): Encoder_linear(
    (activation): ReLU()
    (encoder_net): Sequential(
      (0): Linear(in_features=100, out_features=64, bias=True)
      (1): ReLU()
      (2): Linear(in_features=64, out_features=64, bias=True)
      (3): ReLU()
      (4): Dropout(p=0.2, inplace=False)
      (5): Linear(in_features=64, out_features=64, bias=True)
      (6): ReLU()
      (7): Dropout(p=0.2, inplace=False)
      (8): Linear(in_features=64, out_features=6, bias=True)
    )
  )
```

```
  (inverse): Inverse(
    (activation): ReLU()
    (inverse_net): Sequential(
      (0): Linear(in_features=12, out_features=64, bias=True)
      (1): ReLU()
      (2): Linear(in_features=64, out_features=64, bias=True)
      (3): ReLU()
      (4): Dropout(p=0.3, inplace=False)
      (5): Linear(in_features=64, out_features=64, bias=True)
      (6): ReLU()
      (7): Dropout(p=0.3, inplace=False)
      (8): Linear(in_features=64, out_features=64, bias=True)
      (9): ReLU()
      (10): Dropout(p=0.3, inplace=False)
      (11): Linear(in_features=64, out_features=64, bias=True)
      (12): ReLU()
      (13): Dropout(p=0.3, inplace=False)
      (14): Linear(in_features=64, out_features=1, bias=True)
    )
  )
  (density): Density(
    (activation): ReLU()
    (density_net): Sequential(
      (0): Linear(in_features=12, out_features=64, bias=True)
      (1): ReLU()
      (2): Linear(in_features=64, out_features=64, bias=True)
      (3): ReLU()
      (4): Dropout(p=0.3, inplace=False)
      (5): Linear(in_features=64, out_features=64, bias=True)
      (6): ReLU()
      (7): Dropout(p=0.3, inplace=False)
      (8): Linear(in_features=64, out_features=64, bias=True)
      (9): ReLU()
      (10): Dropout(p=0.3, inplace=False)
      (11): Linear(in_features=64, out_features=64, bias=True)
      (12): ReLU()
      (13): Dropout(p=0.3, inplace=False)
      (14): Linear(in_features=64, out_features=1, bias=True)
    )
  )
  (rho): Rho(
    (activation): ReLU()
    (rho_net): Sequential(
      (0): Linear(in_features=6, out_features=64, bias=True)
      (1): ReLU()
      (2): Linear(in_features=64, out_features=64, bias=True)
      (3): ReLU()
      (4): Dropout(p=0.3, inplace=False)
      (5): Linear(in_features=64, out_features=64, bias=True)
      (6): ReLU()
      (7): Dropout(p=0.3, inplace=False)
      (8): Linear(in_features=64, out_features=2, bias=True)
    )
  )
)
```

C Additional Experimental Details

C.1 Reproducibility

We release our code and data on the website at
`https://anonymous.4open.science/r/state-abstraction-588A/README.md`
The hyper-parameters to train the proposed forward and backward abstractions can be found in Table B.1.

C.2 Experimental settings and additional results

For both environments we use Adam Kingma & Ba (2014) optimizer, with learning rate 0.001 in Cartpole and 0.003 in LunarLander. Model architectures and hyper-parameters are outlined in B. When conducting OPE, the FQE network has 3 hidden layers with 64 nodes per hidden layer for abstraction methods, and is equipped with 5 hidden layers with 128 nodes per hidden layer for non-abstracted observations (shown as 'FQE' in the plot).

C.2.1 CartPole-v0

Data generating processes

We manually insert 296 irrelevant features in the state, each following a first order auto-regressive model (AR(1))

$$\mathbb{P}(S_{t+1,j}|S_t, A_t) = \mathbb{P}(S_{t+1,j}|S_{t,j}), \quad j = 5, \dots, 300.$$

We also define a new state-action-dependent reward as

$$\mathcal{R}(s_t, a_t) = 1 - 2s_{t,1}^2 - 5s_{t,3}^2,$$

where $s_{t,1}$ and $s_{t,3}$ are the first feature (cart position) and third feature (pole angle) of the state s_t, to replace the original constant rewards. The number of trajectories n in the offline dataset is chosen from $\{5, 8, 15, 30\}$, where each trajectory contains approximately 40 decision points. The target policy is determined by the pole angle: we push the cart to the left if the angle is negative and to the right if it is positive. Namely,

$$\pi(s_t) = \mathbb{1}(s_{t,3} > 0).$$

The behavior policy that generates the batch data is set to an ϵ-greedy policy with respect to the target policy, with $\epsilon \in \{0.1,\ 0.3,\ 0.5,\ 0.7\}$. Results are averaged over 30 runs for each (n, ϵ) pair.

Model parameters

For the proposed forward and backward models, we set the abstracted state dimension as 100. For the two-step method, we apply backward abstraction followed by forward abstraction, reducing the dimension from $300 \to 100 \to 6$ for $\epsilon \in \{0.1, 0.3\}$. We change the abstracted dimension to $300 \to 100 \to 2$ for $\epsilon \in \{0.5, 0.7\}$.

C.2.2 LunarLander-v2

Data generating processes

We similarly insert 292 irrelevant auto-regressive features in the state:

$$\mathbb{P}(S_{t+1,j}|S_t, A_t) = \mathbb{P}(S_{t+1,j}|S_{t,j}), \quad j = 9, \dots, 300.$$

The number of trajectories n in the offline dataset is chosen from $\{7, 13, 20\}$, where trajectory length differs significantly in this environment. Some lengthy episodes can have length larger than 100000 while short episodes have fewer than 100 decision points. When trained and evaluated on the short episodes, OPE methods will fail due to huge distributional drift. We therefore truncate the episode length at 1000 if it exceeds, define it as long episode and those fewer than 1000 as short episodes. When generating trajectories, we use a long-short combination for each size: $\{7 = 5_{long} + 2_{short}, 13 = 10_{long} + 3_{short}, 20 = 15_{long} + 5_{short}\}$. The target policy is an estimated optimal policy pre-trained by an DQN agent whereas the behavior policy again ϵ-greedy to the target policy with $\epsilon \in \{0.1,\ 0.3,\ 0.5\}$. Results are averaged over 30 runs for each (n, ϵ) pair and are reported in Figure C.1

Model parameters

For forward and backward models, we abstract the original state dimension from $300 \rightarrow 100$, and for two-step method we reduce dimensions from $300 \rightarrow 50 \rightarrow 4$, by first using forward model and then backward model.

Pre-trained agent

We pre-train an agent by using DQN as our target policy. The agent is trained until there exists an episode that has accumulative discounted rewards exceeding 200 with discounted rate $\gamma = 0.99$. We evaluated oracle value (61.7) of the optimized agent by Monte Carlo method with the same discounted rate. The agent model architecture is as follow:

```
    DQN(
  (fc1): Linear(in_features=8, out_features=64, bias=True)
  (fc2): Linear(in_features=64, out_features=64, bias=True)
  (fc3): Linear(in_features=64, out_features=4, bias=True)
)
```

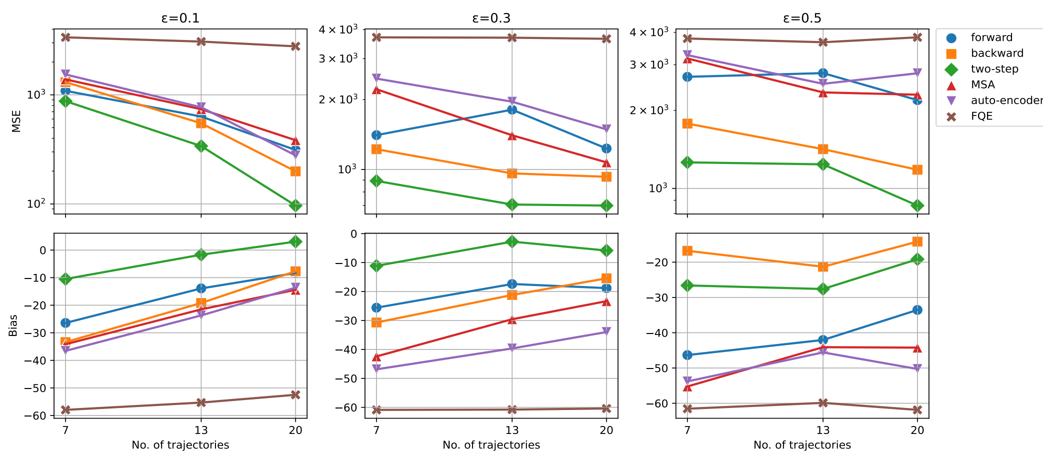


Figure C.1: MSEs and biases of FQE estimators when applied to ground and abstract state spaces with various abstractions. The behavior policy is ϵ-greedy with respect to the target policy, with $\epsilon = 0.1, 0.3, 0.5$ from left to right.

C.3 Licences for existing assets

We consider two environments from OpenAI Gym (Brockman et al., 2016), "CartPole-v0" and "LunarLander-v2" with the MIT License and Copyright (c) 2016 OpenAI (`https://openai.com`).

C.4 Computing resources

C.4.1 CartPole-v0

To build Figure 4, we trained 3 abstraction methods and one non-abstraction method on 4 different sizes of data, each with 30 runs, under 4 values. Each run take approximately 1.5 minutes for four methods on a E2-series CPU with 64GB memory on Google Cloud Platform (GCP). It takes about 12 compute hours to complete all the experiments in the figure.

C.4.2 LunarLander-v2

To build Figure C.1, we trained 3 abstraction methods and one non-abstraction method on 3 different sizes of data, each with 30 runs, under 3 values. In average, each run takes approximately 4 minutes for four methods on a E2-series CPU with 64GB memory on GCP. It takes about 18 compute hours to complete all the experiments in the figure.

D Limitations

Our proposal presents several limitations. Firstly, although empirical results validate the effectiveness of the proposed state abstraction for OPE, we have not conducted a theoretical analysis to determine if state abstraction leads to a more efficient OPE estimator with reduced MSE compared to estimators without abstraction. Additionally, we have not theoretically examined if the two-step procedure's estimator achieves a smaller MSE than estimators derived from single-iteration forward or backward abstraction. We leave these aspects for future research.

E Technical proofs

We provide the detailed proofs of our theorems (Theorems 1, 2, 3, 4) in this section.

Notations. For events or random variables A, B, C, $A \perp\!\!\!\perp B$ means the independence between A and B whereas $A \perp\!\!\!\perp B|C$ means the conditional independence between A and B given C.

E.1 Proof of Theorem 1

We prove Theorem 1 in this subsection. We first prove under Q^π-, ρ^π- or w^π-irrelevance, the corresponding methods remain valid when applied to the abstract state space:

- Q^π**-irrelevance**. By definition, Q^π is the expected return given an initial state S_1 and A_1. Under Q^π-irrelevance, the Q-function depends on S_1 only through $\phi(S_1)$. It follows that Q^π equals the expected return given $\phi(S_1)$ and A_1, the latter being Q_ϕ^π – the Q-function when restricted to the abstract state space, i.e., $Q_\phi^\pi(a, \phi(s)) = \sum_{t\geq 1} \gamma^{t-1} \mathbb{E}^\pi[R_t | A_1 = a, \phi(S_1) = \phi(s)]$. It follows that

$$\begin{aligned}\mathbb{E}[f_1(Q^\pi)] =& \sum_{a,s} \pi(a|s) Q^\pi(a,s) \mathbb{P}(S_1 = s)\\ =& \sum_{a,s} \pi(a|s) Q_\phi^\pi(a, \phi(s)) \mathbb{P}(S_1 = s)\\ =& \mathbb{E}[f_1(Q_\phi^\pi)].\end{aligned}$$

- ρ^π**-irrelevance**. We first establish the equivalence between ρ^π and ρ_ϕ^π – the IS ratio defined on the abstract state space. Under ρ^π-irrelevance, $\rho^\pi(a, s)$ becomes a constant function of $x = \phi(s)$. Consequently, for any conditional probability mass function (pmf) $f(s|x)$ such that $\sum_{s\in\phi^{-1}(x)} f(s|x) = 1$, we have $\rho^\pi(a, s) = \sum_{s\in\phi^{-1}(x)} f(s|x)\rho^\pi(a, s)$. By setting $f(s|x)$ to the pmf of $S_t = s$ given $A_t = a$ and $\phi(S) = x$, it follows that

$$\rho^\pi(a, s) = \sum_{s\in\phi^{-1}(x)} \mathbb{P}(S_t = s | A_t = a, \phi(S_t) = x) \rho^\pi(a, s). \tag{E.1}$$

  Notice that

$$\mathbb{P}(S_t = s | A_t = a, \phi(S_t) = x) = \frac{\mathbb{P}(A_t = a, S_t = s | \phi(S_t) = x)}{\mathbb{P}(A_t = a | \phi(S_t) = x)}.$$

  The denominator equals $b_{\phi,t}(a|x)$, the behavior policy when restricted to the abstract state space at time t. Notice that this behavior policy can be non-stationary over time, despite that b being time-invariant. As for the numerator, it is straightforward to show that it equals $b(a|s)\mathbb{P}(S_t = s|\phi(S_t) = x)$. This together with (E.1) yields

$$\rho^\pi(a, s) = \sum_{s\in\phi^{-1}(x)} \frac{\pi(a|s)}{b_{\phi,t}(a|x)} \mathbb{P}(S_t = s|\phi(S_t) = x) = \frac{\pi_{\phi,t}(a|x)}{b_{\phi,t}(a|x)}, \tag{E.2}$$

  where $\pi_{\phi,t}$ denotes the target policy confined on the abstract state space at time t. The last term in (E.2) is given by $\rho_{\phi,t}^\pi$. Consequently, the cumulative IS ratio $\rho_{1:t}^\pi$ is equal to $\prod_{k=1}^t \rho_{\phi,k}^\pi(A_k, \phi(S_k))$. This in turn yields $\mathbb{E}[f_2(\rho^\pi)] = \mathbb{E}[f_2(\rho_\phi^\pi)]$.

- w^π**-irrelevance**. Similar to the proof under ρ^π-irrelevance, the key lies in establishing the equivalence between $w^\pi(a,s)$ and $w^\pi_\phi(a,\phi(s))$, the latter being the MIS ratio defined on the abstract state space. Once this has been proven, it is immediate to see that $\mathbb{E}[f_3(w^\pi)] = \mathbb{E}[f_3(w^\pi_\phi)]$, so that MIS remains valid when applied to the abstract state space.

  As discussed in Section 2.3, to guarantee the unbiasedness of the MIS estimator, we additionally require a stationarity assumption. Under this requirement, for a given state-action pair (S,A) in the offline data, its joint pmf function can be represented as $p_\infty \times b$ where p_∞ denotes the marginal state distribution under the behavior policy. Additionally, let p_t^π denote the pmf of S_t generated under the target policy π. The MIS ratio can be represented by

$$w^\pi(a,s) = \frac{(1-\gamma)\sum_{t\geq 1}\gamma^{t-1}p_t^\pi(s)\pi(a|s)}{p_\infty(s)b(a|s)}.$$

  Similar to (E.2), under w^π-irreleavance, it follows that

$$\begin{aligned} w^\pi(a,s) &= (1-\gamma)\sum_{s\in\phi^{-1}(x)}\frac{\sum_{t\geq 1}\gamma^{t-1}p_t^\pi(s)\pi(a|s)}{p_\infty(s)b_\phi(a|x)}\mathbb{P}(S=s|\phi(S)=x) \\ &= \frac{(1-\gamma)\sum_{s\in\phi^{-1}(x)}\sum_{t\geq 1}\gamma^{t-1}p_t^\pi(s)\pi(a|s)}{p_\infty(x)b_\phi(a|x)}. \end{aligned}$$

  Here, the subscript t in b_ϕ and S is dropped due to stationarity. Additionally, $p_\infty(x)$ is used to denote the probability mass function (pmf) of $\phi(S)$, albeit with a slight abuse of notation. Moreover, the numerator represents the discounted visitation probability of $(A,\phi(S))$ under π. This proves that $w^\pi(a,s) = w^\pi_\phi(a,\phi(s))$.

Finally, we establish the validity of DRL. According to the doubly robustness property, DRL is valid when either Q^π or w^π is correctly specified. Under Q^π-irrelevance, we have $Q^\pi(a,s) = Q^\pi_\phi(a,\phi(s))$ and thus DRL remains valid when applied to the abstract state space. Similarly, we have $w^\pi(a,s) = w^\pi_\phi(a,\phi(s))$ under w^π-irrelevance, which in turn implies DRL's validity. This completes the proof.

E.2 Proof of Theorem 2

We prove Theorem 2 in this subsection.

- For any $s^{(1)}$ and $s^{(2)}$ satisfies (2), we aim to prove

$$Q^\pi(a,s^{(1)}) = Q^\pi(a,s^{(2)}).$$

  Toward that end, we use the induction method. Denote

$$\begin{aligned} Q_j^\pi(a,s) &= \mathbb{E}^\pi\left[\sum_{t=1}^{j}\gamma^{t-1}R_t|S_1=s,A_1=a\right], \text{ and} \\ V_j^\pi(s) &= \mathbb{E}^\pi\left[\sum_{t=1}^{j}\gamma^{t-1}R_t|S_1=s\right]. \end{aligned}$$

  Under reward-irrelevance, we have

$$\begin{aligned} Q_1^\pi(a,s^{(1)}) =&\mathbb{E}^\pi\left[R_1|S_1=s^{(1)},A_1=a\right] \\ =&\mathcal{R}(a,s^{(1)}) \\ =&\mathcal{R}(a,s^{(2)}) \\ =&Q_1^\pi(a,s^{(2)}). \end{aligned}$$

Together with π-irrelevance, we obtain that

$$\begin{aligned}
V_1^\pi(s^{(1)}) =&\mathbb{E}^\pi\left[R_1|S_1=s^{(1)}, A_1=a\right]\pi(a|s^{(1)})\\
=&\mathcal{R}(a,s^{(1)})\pi(a|s^{(1)})\\
=&\underbrace{\mathcal{R}(a,s^{(2)})}_{\text{reward-irrelevant}}\underbrace{\pi(a|s^{(2)})}_{\pi-\text{irrelevant}}\\
=&V_1^\pi(s^{(2)}).
\end{aligned}$$

Suppose we have shown that the following holds for any $j<T$,

$$Q_j^\pi(a,s^{(1)})=Q_j^\pi(a,s^{(2)}) \text{ and } V_j^\pi(s^{(1)})=V_j^\pi(s^{(2)}). \tag{E.3}$$

Our goal is to show (E.3) holds with $j=T$.

We similarly define $Q_{j,\phi}^\pi$ and $V_{j,\phi}^\pi$ as the Q- and value functions defined on the abstract state space. Similar to the proof of Theorem 1, we can show that $Q_j^\pi=Q_{j,\phi}^\pi$ and $V_j^\pi=V_{j,\phi}^\pi$ for any $j<T$. It follows that

$$\begin{aligned}
Q_T^\pi(a,s^{(1)}) =&\mathbb{E}^\pi\left[\sum_{t=1}^T\gamma^{t-1}R_t|S_1=s^{(1)},A_1=a\right]\\
=&\mathbb{E}^\pi\left[\sum_{t=2}^T\gamma^{t-1}R_t|S_1=s^{(1)},A_1=a\right]+\mathcal{R}(a,s^{(1)})\\
=&\gamma\mathbb{E}^\pi\sum_{s'\in\mathcal{S}}\left[\sum_{t=2}^T\gamma^{t-1}R_t|S_2=s'\right]\mathcal{T}(s'|s^{(1)},a)+\mathcal{R}(a,s^{(1)})\\
=&\gamma\mathbb{E}^\pi\sum_{x'\in\mathcal{X}}\sum_{s'\in\phi^{-1}(x')}\left[\sum_{t=2}^T\gamma^{t-2}R_t|S_2=s'\right]\mathcal{T}(s'|s^{(1)},a)+\mathcal{R}(a,s^{(1)})\\
=&\gamma\sum_{x'\in\mathcal{X}}\sum_{s'\in\phi^{-1}(x')}V_{T-1}^\pi(s')\mathcal{T}(s'|s^{(1)},a)+\mathcal{R}(a,s^{(1)})\\
=&\gamma\sum_{x'\in\mathcal{X}}\underbrace{V_{T-1,\phi}^\pi(x')}_{\text{by (E.3)}}\sum_{s'\in\phi^{-1}(x')}\mathcal{T}(s'|s^{(1)},a)+\mathcal{R}(a,s^{(1)})\\
=&\gamma\sum_{x'\in\mathcal{X}}\underbrace{V_{T-1,\phi}^\pi(x')}_{\text{by (E.3)}}\underbrace{\sum_{s'\in\phi^{-1}(x')}\mathcal{T}(s'|s^{(2)},a)}_{(2)}+\mathcal{R}(a,s^{(2)})\\
=&Q_T^\pi(a,s^{(2)}).
\end{aligned}$$

This together with π-irrelevance proves V_T^π-irrelevance. Consequently, (E.3) holds for any $j\geq 1$. Since $Q_j^\pi\to Q^\pi$ as $j\to\infty$, we obtain Q^π-irrelevance.

- We will prove that the MIS estimator constructed on the abstract state space remains valid. With a slight abuse of notation, we use $p_t^\pi(a,x)$ to denote the probability $\mathbb{P}^\pi(A_t=a,\phi(S_t)=x)$. Under the stationarity assumption, direct calculations yield

$$
\begin{aligned}
\mathbb{E}[f_3(w_\phi^\pi)] =& \mathbb{E}\left[(1-\gamma)^{-1} w_\phi^\pi(A, \phi(S)) R\right] \\
=& \mathbb{E}\left[(1-\gamma)^{-1} w_\phi^\pi(A, \phi(S)) \mathcal{R}\big(A, S\big)\right] \\
=& \mathbb{E}\left[(1-\gamma)^{-1} w_\phi^\pi(A, \phi(S)) \quad \underbrace{\mathcal{R}\big(A, \phi(S)\big)}_{\text{reward-irrelevant}}\right] \\
=& \sum_{a\in\mathcal{A}, x\in\mathcal{X}} \sum_{t=1}^{+\infty} \gamma^{t-1} p_t^\pi(a, x) \mathcal{R}_\phi(a, x) \\
=& \sum_{a\in\mathcal{A}, x\in\mathcal{X}} \sum_{s\in\phi^{-1}(x)} \sum_{t=1}^{+\infty} \gamma^{t-1} \pi(a|s) p_t^\pi(s) \mathcal{R}(a, s) \\
=& \sum_{t=1}^{+\infty} \gamma^{t-1} \mathbb{E}^\pi(R_t) \\
=& \mathbb{E}[f_3(w^\pi)]
\end{aligned}
$$

Notice that we only require reward-irrelevance in the above proof.

- It suffices to show that

$$
\mathbb{E}[\rho_{1:t}^\pi R_t] = \mathbb{E}[\prod_{k=1}^t \rho_{\phi,t}^\pi (A_k, \phi(S_k)) R_t], \tag{E.4}
$$

for any t. Under the Markov assumption, R_t is independent of past state-action pairs given A_t and S_t. Consequently, the left-hand-side can be represented as

$$
\mathbb{E}[\mathbb{E}(\rho_{1:t-1}^\pi | A_t, S_t) \rho^\pi(A_t, S_t) R_t].
$$

Additionally, since the generation A_t depends only on S_t, the inner expectation equals $\mathbb{E}(\rho_{1:t-1}^\pi | S_t)$ which can be further shown to equal to $p_t^\pi(S_t)/p_\infty(S_t)$. This allows us to represent the left-hand-side of (E.4) by

$$
\mathbb{E}\Big[\frac{p_t^\pi(S_t)}{p_\infty(S_t)} \rho^\pi(A_t, S_t) R_t\Big]. \tag{E.5}
$$

Using similar arguments in proving the validity of MIS estimator, under reward-irrelevance, (E.5) can be shown to equal to

$$
\sum_{a\in\mathcal{A}, x\in\mathcal{X}} p_t^\pi(a, x) \mathcal{R}_\phi(a, x). \tag{E.6}
$$

Under transition-irrelevance, the data triplets $(\phi(S), A, R)$ forms an MDP, satisfying the Markov assumption. Let $\mathcal{T}_\phi$ denote the resulting transition function. Together with π-irrelevance, we can rewrite (E.6) as

$$
\sum_{\substack{a_1,\cdots,a_t\in\mathcal{A} \\ x_1,\cdots,x_t\in\mathcal{X}}} \rho_0(x_1) \prod_{k=1}^{t-1} \Big[\pi_\phi(a_k|x_k) \mathcal{T}_\phi(x_{k+1}|a_k, x_k)\Big] \pi_\phi(a|x_t) \mathcal{R}_\phi(a, x).
$$

Notice that $\mathcal{T}_\phi$ is independent of the target policy π. Using the change of measure theorem, we can represent above expression by $\mathbb{E}(\rho_{1:t,\phi}^\pi R_t)$ where $\rho_{1:t,\phi}^\pi$ denotes the cumulative IS ratio defined on the abstract state space. This completes the proof.

- Since model-irrelvance implies Q^π-irrelevance, the conclusion directly follows from the last conclusion of Theorem 1.

E.3 Proof of Theorem 3

At the begging of the proof, we name the phenomena as the Inverse Markovianity, namely the reversed state-action pairs maintain the Markov property.

- ρ^π-irrelevance directly follows from the definition of backward-model-irrelevance. To show w^π-irrelevance, we divide the proof into two steps.
  (1) In the first step, we will prove that if ϕ satisfies the backward-model-irrelevance, then
  $$\rho^\pi(A_{t-k}, S_{t-k}) \perp\!\!\!\perp S_t|\phi(S_t), 1 \le k \le t-1. \tag{E.7}$$
  It follows from equation (3) that
  $$\mathbb{P}\big(\phi(S_{t-k}) = x|S_{t-k+1}\big) = \mathbb{P}\big(\phi(S_{t-k}) = x|\phi(S_{t-k+1})\big), 1 \le k \le t-1.$$
  We can use the induction method to prove that for $1 \le k \le t-1$,
  $$\rho^\pi(A_{t-k}, S_{t-k}) \perp\!\!\!\perp S_t|\phi(S_t). \tag{E.8}$$
  For $k = 1$, we have for any positive constant c,
  $$\begin{aligned}\mathbb{P}\big(\rho^\pi(A_{t-1}, S_{t-1}) = c|S_t\big) =& \mathbb{P}[\rho^\pi_{\phi,t-1}\big(A_{t-1}, \phi(S_{t-1})\big) = c|S_t] \\ =& \mathbb{P}[\rho^\pi_{\phi,t-1}\big(A_{t-1}, \phi(S_{t-1})\big) = c|\phi(S_t)],\end{aligned} \tag{E.9}$$
  where the first equation is due to ρ^π-irrelevance and the second equation follows from (3). This yields
  $$\rho^\pi(A_{t-1}, S_{t-1}) \perp\!\!\!\perp S_t|\phi(S_t).$$
  We assume that for $k \le t-2$ the formulation (E.8) holds. Now, we prove that for $k = t-1$, (E.8) successes. By similar arguments to that of (E.9), we get
  $$\begin{aligned}\mathbb{P}\big(\rho^\pi(A_1, S_1) = c|S_t\big) =& \mathbb{P}[\mathbb{P}\big(\rho^\pi(A_1, S_1) = c|S_2, A_2, S_t, A_t\big)|S_t] \\ =& \mathbb{P}[\mathbb{P}\big(\rho^\pi(A_1, S_1) = c|S_2\big)|S_t] \\ =& \mathbb{P}[g\big(\phi(S_2)\big)|S_t].\end{aligned} \tag{E.10}$$
  To prove this, we need to show that for any $1 \le k \le t-1$, we have
  $$\mathbb{P}\big(\phi(S_{t-k}) = x|S_t\big) = \mathbb{P}\big(\phi(S_{t-k}) = x|\phi(S_t)\big). \tag{E.11}$$
  The definition of inverse model implies when $k = 1$, (E.11) successes. We assume that for $k \le t-2$ the formulation (E.11) successes. Now, we prove that for $k = t-1$, (E.11) also hold.
  $$\begin{aligned}\mathbb{P}\big(\phi(S_1) = x|S_t\big) =& \mathbb{P}[\mathbb{P}\big(\phi(S_1) = x|S_2, S_t\big)|S_t] \\ =& \underbrace{\mathbb{P}[\mathbb{P}\big(\phi(S_1) = x|S_2\big)|S_t]}_{\text{Inverse Markovianity}} \\ =& \underbrace{\mathbb{P}[\mathbb{P}\big(\phi(S_1) = x|\phi(S_2)\big)|S_t]}_{\text{Inverse Markovianity}} \\ =& \underbrace{\mathbb{P}[g\big(\phi(S_2)\big)|S_t]}_{\text{Inverse Markovianity}} \\ =& \underbrace{\mathbb{P}[g\big(\phi(S_2)\big)|\phi(S_t)]}_{\text{(E.11)}}.\end{aligned}$$
  This proves (E.11). Combing (E.10) and (E.11), we can get
  $$\mathbb{P}\big(\rho^\pi(A_1, S_1) = c|S_t\big) = \mathbb{P}[g\big(\phi(S_2)\big)|\phi(S_t)].$$
  Then we prove (E.7).
  (2)In the second step, we will prove that if ϕ satisfies equation (E.7) and ρ^π-irrelevance, it is w^π-irrelevant, namely for any $s^{(1)}$ and $s^{(2)}$ satisfying $\rho^\pi_t(a, s^{(1)}) = \rho^\pi_t(a, s^{(2)})$, they will satisfy
  $$w^\pi(a, s^{(1)}) = w^\pi(a, s^{(2)}).$$

It follows from the definition of state abstraction, $s^{(1)}$ and $s^{(2)}$, we have

$$\mathbb{P}(X_t|S_t = s^{(1)}) = \mathbf{1}\big(s^{(1)} \in \phi^{-1}(X_t)\big) = \mathbf{1}\big(s^{(2)} \in \phi^{-1}(X_t)\big) = \mathbb{P}(X_t|S_t = s^{(2)}). \quad \text{(E.12)}$$

By (E.12) and (E.7), we have

$$\begin{aligned}
w^\pi(a, s^{(1)}) =& \frac{(1-\gamma)\sum_{t=1}^T \gamma^{t-1}\mathbb{P}^\pi(A_t = a, S_t = s^{(1)})}{\mathbb{P}(A = a, S = s^{(1)})} \\
=& \frac{(1-\gamma)\sum_{t=1}^T \gamma^{t-1}\mathbb{P}^\pi(A_t = a|S_t = s^{(1)})\mathbb{P}^\pi(S_t = s^{(1)})}{\mathbb{P}(A = a|S = s^{(1)})\mathbb{P}^b(S = s^{(1)})} \\
=& \frac{(1-\gamma)\sum_{t=1}^T \gamma^{t-1}\rho_t^\pi(a, s^{(1)})\mathbb{P}^\pi(S_t = s^{(1)})}{\mathbb{P}^b(S = s^{(1)})} \\
=& \frac{(1-\gamma)\sum_{t=1}^T \gamma^{t-1}\rho_t^\pi(a, s^{(1)})\mathbb{E}^\pi[\mathbf{1}(S_t = s^{(1)})]}{\mathbb{E}^b[\mathbf{1}(S_t = s^1)]} \\
=& \frac{(1-\gamma)\sum_{t=1}^T \gamma^{t-1}\rho_t^\pi(a, s^{(1)})\mathbb{E}^b[\mathbf{1}(S_t = s^{(1)})\prod_{j=1}^{t-1}\rho_j^\pi(A_j, S_j)]}{\mathbb{E}^b[\mathbf{1}(S_t = s^{(1)})]} \\
=& \frac{(1-\gamma)\sum_{t=1}^T \gamma^{t-1}\rho_t^\pi(a, s^{(1)})\mathbb{E}^b\left[\mathbb{E}^b\left(\mathbf{1}(S_t = s^{(1)})\prod_{j=1}^{t-1}\rho_j^\pi(A_j, S_j)|X_t\right)\right]}{\mathbb{E}^b[\mathbf{1}(S_t = s^{(1)})]} \\
=& \underbrace{\frac{(1-\gamma)\sum_{t=1}^T \gamma^{t-1}\rho_t^\pi(a, s^{(1)})\mathbb{E}^b\left[\mathbb{E}^b\left(\mathbf{1}(S_t = s^{(1)})|X_t\right)\mathbb{E}^b\left(\prod_{j=1}^{t-1}\rho_j^\pi(A_j, S_j)|X_t\right)\right]}{\mathbb{E}^b[\mathbf{1}(S_t = s^{(1)})]}}_{\text{by (E.7)}} \\
=& (1-\gamma)\sum_{t=1}^T \gamma^{t-1}\rho_t^\pi(a, s^{(1)})\mathbb{E}^b\left(\frac{\mathbb{P}(X_t|S_t = s^{(1)})\prod_{j=1}^{t-1}\rho_j^\pi(A_j, S_j)}{\mathbb{P}(X_t)}\right) \\
=& \underbrace{(1-\gamma)\sum_{t=1}^T \gamma^{t-1}\rho_t^\pi(a, s^{(2)})\mathbb{E}^b\left(\frac{\mathbb{P}(X_t|S_t = s^{(2)})\prod_{j=1}^{t-1}\rho_j^\pi(A_j, S_j)}{\mathbb{P}(X_t)}\right)}_{\text{by (E.12)}} \\
=& w^\pi(a, s^{(2)}).
\end{aligned}$$

Then, we can conclude that backward-model-irrelevance implies the ρ^π-irrelevance and w^π-irrelevance.

- It follows from the definition of Q-function-based method that

$$\begin{aligned}
\mathbb{E}[f_1(Q_\phi^\pi)] =& \sum_{a,x} Q_\phi^\pi(a, x)\pi(a|x)\mathbb{P}(\phi(S_1) = x) \\
=& \sum_{a,x}\mathbb{E}^\pi\Big[\sum_{t=1}^{+\infty}\gamma^{t-1}R_t|X_1 = x, A_1 = a\Big]\pi(a|x)\mathbb{P}(X_1 = x) \\
=& \sum_{a,x,r}\sum_{t=1}^{+\infty}\gamma^{t-1}r\mathbb{P}^\pi\Big[r|X_1 = x, A_1 = a\Big]\pi(a|x)\mathbb{P}(X_1 = x) \\
=& \mathbb{E}^\pi\Big[\sum_{t=1}^{+\infty}\gamma^{t-1}R_t\Big] \\
=& \mathbb{E}[f_1(Q^\pi)].
\end{aligned}$$

- The conclusion directly follows from the last conclusion of Theorem 1, and the first conclusion of Theorem 3.

E.4 Proof of Theorem 4

Theorem 4 directly follows from Theorem 2 and Theorem 3. We just list the Q-function based method and initialization from forward state abstraction. Firstly, based on the first conclusions in

Theorems 1 and 2, we can get that Q-function based method remains valid. Namely, for the forward state abstraction function ϕ_1, we have

$$\mathbb{E}[f_1(Q^{\pi}_{\phi_1})] = \mathbb{E}[f_1(Q^{\pi})].$$

Based on $\phi_1(\mathcal{S}) = \mathcal{X}_1$, we derive the backward state abstraction ϕ_2. The second conclusion in Theorem 3 indicates

$$\mathbb{E}[f_1(Q^{\pi}_{\phi_2 \circ \phi_1})] = \mathbb{E}[f_1(Q^{\pi}_{\phi_1})] = \mathbb{E}[f_1(Q^{\pi})].$$

This indicates that after one step of the forward-backward iteration, the Q-value-based function still works.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We confirm the claims in abstract and introduction are accurately reflected by methodology, theory, and experiments in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss the limitations of the work in Appendix D.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are provided in Section 3.1 and Section 3.2. The proof is attached into Appendix E.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The detailed setting for reproducibility is provided in Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The dataset and code in this paper are either publicly available or submitted as a new asset, see Section C.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The hyper-parameters are given in Table B.1 in Appendix B. The details for experiments are shown in Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The results are accompanied by the error bars as can be seen from Figure 4 in the main text Section 4 and Figure C.1 in the Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C.4 provides compute resources for experiments.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This paper aims to contribute to the advancement of Machine Learning. Although there are potential societal impacts stemming from our research, we believe none require specific emphasis in this context.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This research does not involve data or models that have a high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: In Appendix C.3, we have explicitly credited the assets used in the paper and explicitly mentioned the corresponding licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented, and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new asset is the implementation of the methods introduced in the paper, see Section 4 and Appendix C. The documentation for the new asset is provided alongside the code in Appendix C.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This research does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.