Scaling Marginalized Importance Sampling to High-Dimensional State-Spaces via State Abstraction

Brahma S. Pavse and Josiah P. Hanna
Department of Computer Science
University of Wisconsin – Madison, Madison, WI
pavse@wisc.edu, jphanna@cs.wisc.edu

Abstract

We consider the problem of off-policy evaluation (OPE) in reinforcement learning (RL), where the goal is to estimate the performance of an evaluation policy, π_e, using a fixed dataset, $\mathcal{D}$, collected by one or more policies that may be different from π_e. Current OPE algorithms may produce poor OPE estimates under policy distribution shift i.e., when the probability of a particular state-action pair occurring under π_e is very different from the probability of that same pair occurring in $\mathcal{D}$ Voloshin et al. (2021); Fu et al. (2021). In this work, we propose to improve the accuracy of OPE estimation by projecting the ground state-space into a lower-dimensional state-space using concepts from the state abstraction literature in RL. Specifically, we consider marginalized importance sampling (MIS) OPE algorithms which compute distribution correction ratios to produce their OPE estimate. In the original state-space, these ratios may have high variance which may lead to high variance OPE. However, we prove that in the lower-dimensional abstract state-space the ratios can have lower variance resulting in lower variance OPE. We then highlight the challenges that arise when estimating the abstract ratios from data, identify sufficient conditions to overcome these issues, and present a minimax optimization problem whose solution yields these abstract ratios. Finally, our empirical evaluation on difficult, high-dimensional state-space OPE tasks shows that the abstract ratios can make MIS OPE estimators achieve lower mean-squared error and more robust to hyperparameter tuning than the ground ratios. [1]

1 Introduction

One of the key challenges when applying reinforcement learning (RL) Sutton and Barto (2018) to real-world tasks is the problem of off-policy evaluation (OPE) Fu et al. (2021); Voloshin et al. (2021). The goal of OPE is to evaluate a policy of interest by leveraging offline data generated by possibly different policies. Solving the OPE problem would enable us to estimate the performance of a potentially risky policy without having to actually deploy it. This capability is especially important for sensitive real-world tasks such as healthcare and autonomous driving. The core OPE problem is to produce accurate policy value estimates under policy distribution shift. This problem is particularly difficult on tasks with high-dimensional state-spaces Voloshin et al. (2021); Fu et al. (2021). For example, consider the AntUMaze problem illustrated on the left side of Figure 1. In this task, an ant-like robot with a high-dimensional state representation moves in a U-shaped maze and receives a reward only for reaching a specific 2D coordinate goal location. The state-space of this task includes information such as 2D location, ant limb angles, torso orientation etc., resulting in a 29-dimensional state-space. The OPE task is to evaluate the performance of a particular policy's ability to take the

[1] This paper will also be published at the Association for the Advancement of Artificial Intelligence (AAAI) 2023.

Offline Reinforcement Learning Workshop at Neural Information Processing Systems, 2022.

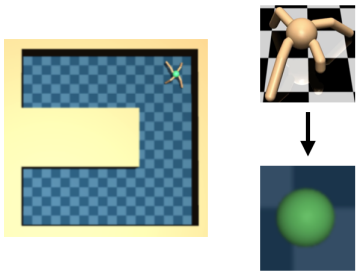


Figure 1: Left side: AntUMaze domain. Right side: Projecting high-dimensional ant into lower-dimensional point-mass.

ant to the 2D goal location using data that may be collected by different policies. Policy distribution shift is common in this type of high-dimensional task since the chances of different policies inducing similar limb angles, torso orientations, paths traversed etc. are incredibly slim. Notice, however, that while different policies may induce different body configurations, they may traverse similar 2D paths since all (successful) policies must move the ant through roughly the same path to reach the goal. Moreover, the only critical information needed from the state-space to determine the ant's per-step reward are its 2D coordinates. Motivated by this observation, we propose to improve the accuracy of OPE algorithms on high-dimensional state-space tasks by projecting the high-dimensional state-space into a lower-dimensional space. This idea is illustrated on the right side of Figure 1 where the ant is reduced to a 2D point-mass.

With this general motivation in mind, in this paper, we leverage concepts from the state abstraction literature Li et al. (2006) to improve the accuracy of marginalized importance sampling (MIS) OPE algorithms which estimate state-action density correction ratios to compute a policy value estimate Liu et al. (2018a); Xie et al. (2019). Due to the low chances of similarity between states of policies in high-dimensional state-spaces, current MIS algorithms can produce high variance state-action density ratios, resulting in high variance OPE estimates. However, if we are given a suitable state abstraction function, we can project the high-dimensional *ground* state-space into a lower-dimensional *abstract* state-space. The projection step increases the chances of similarity between these lower-dimensional states, resulting in low variance density ratios and OPE estimates. To the best of our knowledge, this work is the first to leverage concepts from state abstraction to improve OPE. We make the following contributions:

1. Theoretical analyses showing: (a) the variance of abstract state-action ratios is at most that of ground state-action ratios; and (b) that our abstract MIS OPE estimator is unbiased, strongly consistent, and can have lower variance than its ground equivalent.
2. An algorithm, based on a popular class of MIS algorithms, to estimate the abstract state-action ratios.
3. An empirical analysis of our estimator on a variety of high-dimensional state-space tasks.

2 Preliminaries

In this section, we discuss relevant background information.

2.1 Notation and Problem Setup

We consider an infinite-horizon Markov decision process (MDP), $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, P, \gamma, d_0 \rangle$, where $\mathcal{S}$ is the state-space, $\mathcal{A}$ is the action-space, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \Delta([0, \infty))$ is the reward function, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition dynamics function, $\gamma \in [0, 1)$ is the discount factor, and $d_0 \in \Delta(\mathcal{S})$ is the initial state distribution. The agent acting, according to policy π, in the MDP generates a trajectory: $s_0, a_0, r_0, s_1, a_1, r_1, ...$, where $s_0 \sim d_0$, $a_t \sim \pi(\cdot|s_t)$, $r_t \sim \mathcal{R}(s_t, a_t)$, and $s_{t+1} \sim P(\cdot|s_t, a_t)$ for $t \geq 0$. We define $r(s, a) := \mathbb{E}_{r \sim \mathcal{R}(s,a)}[r]$ and the agent's discounted state-

action occupancy measure under policy π as d_π:

$$d_\pi(s,a) := \lim_{T\to\infty} \frac{\sum_{t=0}^{T-1} \gamma^t d_\pi(s_t,a_t)}{\sum_{t=0}^{T-1}\gamma^t}$$

where $d_\pi(s_t,a_t)$ is the probability the agent will be in state s and take action a at time-step t under policy π.

In the infinite-horizon setting, we define the performance of policy π to be its average reward, $\rho(\pi) := \mathbb{E}_{(s,a)\sim d_\pi, r\sim\mathcal{R}(s,a)}[r]$. Note that $\mathbb{E}_{(s,a)\sim d_\pi, r\sim\mathcal{R}(s,a)}[r] = (1-\gamma)\,\mathbb{E}_{s_0\sim d_0, a_0\sim\pi}[q^\pi(s_0,a_0)]$ where $q^\pi(s,a) := \mathbb{E}[\sum_{k=0}^{\infty}\gamma^t r_{k+t} | s_t = s, a_t = a]$ is the action-value function which satisfies $q^\pi(s,a) = r(s,a) + \gamma\,\mathbb{E}_{s'\sim P(s,a), a'\sim\pi(s')}[q^\pi(s',a')]$.

2.2 Off-Policy Evaluation (OPE)

In behavior-agnostic off-policy evaluation (OPE), the goal is to estimate the performance of an evaluation policy π_e given only a fixed offline data set of transition tuples, $\mathcal{D} := \{(s_i, a_i, s_i', r_i)\}_{i=1}^{mT}$, where $(s_i,a_i)\sim d_\mathcal{D}$, m is the batch size (number of trajectories), and T is the fixed length of each trajectory, generated by *unknown* and possibly *multiple* behavior policies. The difficulty in OPE is to estimate $\rho(\pi_e)$ under d_{π_e} given samples only from $d_\mathcal{D}$.

We define the average-reward in dataset $\mathcal{D}$ to be $\bar{r}_\mathcal{D} := \mathbb{E}_{(s,a)\sim d_\mathcal{D}, r\sim\mathcal{R}(s,a)}[r]$. As in prior OPE work, we assume that if $d_{\pi_e}(s,a) > 0$ then $d_\mathcal{D}(s,a) > 0$. Empirically, we measure the accuracy of an estimate $\hat{\rho}(\pi_e)$ by generating M datasets and then computing the *relative* mean-squared error (MSE): $\text{MSE}(\hat{\rho}(\pi_e)) := \frac{1}{M}\sum_{i=1}^{M}\frac{(\rho(\pi_e)-\hat{\rho}_i(\pi_e))^2}{(\rho(\pi_e)-\bar{r}_{\mathcal{D}_i})^2}$, where $\hat{\rho}_i(\pi_e)$ is computed using dataset $\mathcal{D}_i$ and $\bar{r}_{\mathcal{D}_i}$ is the average reward in $\mathcal{D}_i$.

2.2.1 Marginalized Importance Sampling (MIS)

In this work, we focus on MIS methods, which evaluate π_e by using the ratio between d_{π_e} and $d_\mathcal{D}$. That is, MIS methods evaluate π_e by estimating $\rho(\pi_e) := \mathbb{E}_{(s,a)\sim d_\mathcal{D}, r\sim\mathcal{R}(s,a)}[\zeta(s,a)r]$, where $\zeta(s,a) := \frac{d_{\pi_e}(s,a)}{d_\mathcal{D}(s,a)}$ is the state-action density ratio for state-action pair (s,a) and $d_\pi(s,a) = d_\pi(s)\pi(a|s)$. When the true ζ is known, the empirical estimate of $\rho(\pi_e)$ is:

$$\hat{\rho}(\pi_e) := \frac{1}{N}\sum_{i=1}^{N}\zeta(s_i,a_i)r(s_i,a_i) \tag{1}$$

where N is the number of samples. In practice, however, ζ is unknown and must be estimated.

One set of ζ-estimation algorithms, which have also shown potential for good OPE performance Voloshin et al. (2021), is the DICE family Yang et al. (2020). While there are many variations, the general DICE optimization problem is:

$$\begin{aligned}&\max_{\zeta:\mathcal{S}\times A\to\mathbb{R}}\min_{\nu:\mathcal{S}\times A\to\mathbb{R}} J(\zeta,\nu) := \\ &\mathbb{E}_{(s,a,s')\sim d_\mathcal{D}, a'\sim\pi_e}[\zeta(s,a)(\nu(s,a)-\gamma\nu(s',a'))] - (1-\gamma)\,\mathbb{E}_{s_0\sim d_0, a_0\sim\pi_e}[\nu(s_0,a_0)]\end{aligned} \tag{2}$$

where the solution to the optimization, $\zeta^*(s,a)$, are the true ratios. The estimator we present in Section 3 builds upon the DICE framework.

2.3 State Abstractions

We define a state abstraction function as a mapping $\phi : \mathcal{S} \to \mathcal{S}^\phi$, where $\mathcal{S}$ is called the *ground* state-space and $\mathcal{S}^\phi$ is called the *abstract* state-space. We consider state abstraction functions that partition the ground state-space into disjoint sets.

We can use ϕ to project the original MDP into a new abstract MDP with the same action-space $\mathcal{A}$ and reward and transition dynamics functions defined as:

$$\mathcal{R}^{\phi}(s^{\phi}, a) = \sum_{s \in \phi^{-1}(s^{\phi})} w(s)\mathcal{R}(s, a)$$

$$P^{\phi}(s'^{\phi}|s^{\phi}, a) = \sum_{s \in \phi^{-1}(s^{\phi}), s' \in \phi^{-1}(s'^{\phi})} w(s)P(s'|s, a)$$

where $w : \mathcal{S} \to [0, 1]$ is a ground state weighting function where for each s^{ϕ}, $\sum_{s \in \phi^{-1}(s^{\phi})} w(s) = 1$ Li et al. (2006). Similarly a policy can be transformed into its abstract equivalent as:

$$\pi^{\phi}(a|s^{\phi}) = \sum_{s \in \phi^{-1}(s^{\phi})} w(s)\pi(a|s).$$

We define the following state-weighting function for an arbitrary policy π: $w_{\pi}(s) = \frac{d_{\pi}(s)}{\sum_{s' \in \phi^{-1}(s^{\phi})} d_{\pi}(s')}$ and only consider abstractions that satisfy:

Assumption 1 (Reward distribution equality)**.** *$\forall s_1, s_2 \in \mathcal{S}$ such that $s_1, s_2 \in s^{\phi}$, $\forall a, \mathcal{R}(s_1, a) = \mathcal{R}(s_2, a)$.*

Assumption 1 implies that, regardless of the choice of the state-weighting function, for given action a, $\forall s \in s^{\phi}$, $\mathcal{R}^{\phi}(s^{\phi}, a) = \mathcal{R}(s, a)$ i.e. the reward distribution of an abstract state equals that of the ground states within that abstract state.

3 Abstract Marginalized Importance Sampling

Marginalized IS methods may suffer from high variance in high-dimensional state-spaces. To potentially reduce this high variance, we propose to first use ϕ to project $\mathcal{D}$ into the abstract state-space to obtain: $\mathcal{D}^{\phi} := \{(s^{\phi}, a, r^{\phi}, s'^{\phi})\}$ where $s^{\phi} = \phi(s)$ and $r^{\phi}(s, a) = r(s, a) \forall s \in s^{\phi}$, and then use the following estimator on $\mathcal{D}^{\phi}$ to estimate π_e^{ϕ}:

Definition 1 (Abstract estimator)**.** *We define our estimator of $\rho(\pi_e^{\phi})$ as follows:*

$$\hat{\rho}(\pi_e^{\phi}) := \frac{1}{N} \sum_{i=1}^{N} \frac{d_{\pi_e^{\phi}}(s_i^{\phi}, a_i)}{d_{\mathcal{D}^{\phi}}(s_i^{\phi}, a_i)} r^{\phi}(s_i^{\phi}, a_i) \tag{3}$$

where N is the number of samples, $d_{\pi^{\phi}}(s^{\phi}, a) = d_{\pi^{\phi}}(s^{\phi})\pi^{\phi}(a|s^{\phi})$ with $d_{\pi^{\phi}}(s^{\phi}) = \sum_{s \in \phi^{-1}(s^{\phi})} d_{\pi}(s)$ and π^{ϕ} constructed using w_{π}.

In the remainder of this section, we first give an example to build intuition for why the abstract ratios $\frac{d_{\pi_e^{\phi}}(s^{\phi}, a)}{d_{\mathcal{D}^{\phi}}(s^{\phi}, a)}$ can be lower variance than the ground ratios and then show theoretically that the OPE estimator given in Equation (3) is strongly consistent and can produce lower variance OPE estimates of $\rho(\pi_e)$ than the ground equivalent. Finally, we identify sufficient conditions under which accurate estimation is possible and adapt an existing DICE algorithm to estimate the abstract ratios.

3.1 A Hard Example for Ground MIS Ratios

We present a hard example for ground MIS ratios in Figure 2 that provides intuition for why the abstract MIS ratios can have lower variance ratios than the ground ratios. Consider two symmetric policies, π_e and $\pi_{\mathcal{D}}$, executed in the ground MDP (left side). In this example, the high variance of the true state-action density ratios $\frac{d_{\pi_e}(s_1, a_0)}{d_{\pi_{\mathcal{D}}}(s_1, a_0)} \approx 0$ and $\frac{d_{\pi_e}(s_2, a_0)}{d_{\pi_{\mathcal{D}}}(s_2, a_0)} \approx 100$ can lead to high variance estimates of $\rho(\pi_e)$. Notice, however, that states s_1 and s_2 are essentially equivalent i.e. $r(s_1, a) = r(s_2, a) \forall a \in \mathcal{A}$ and can be aggregated together into a single state, s_1^{ϕ} (Assumption 1). We find that the state-action density ratio in this abstract MDP (right side) $\frac{d_{\pi_e^{\phi}}(s_1^{\phi}, a_0)}{d_{\pi_{\mathcal{D}}^{\phi}}(s_1^{\phi}, a_0)} = 1$ is of low variance, which can lead to low variance estimates of $\rho(\pi_e)$. In fact, with Theorem 1 we prove (proof in Appendix A.3) that the abstract ratios are guaranteed to have variance at most that of the ground ratios:

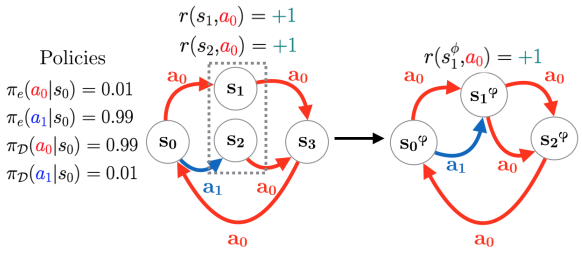


Figure 2: TwoPath MDP where ground density ratios for (s_1, a_0) and (s_2, a_0) are high variance. However, upon aggregation of equivalent states (grey dotted lines), the abstract density ratio of (s_1^ϕ, a_0) is low variance.

Theorem 1. $Var\left(\frac{d_{\pi_e^\phi}(s^\phi,a)}{d_{\mathcal{D}^\phi}(s^\phi,a)}\right) \leq Var\left(\frac{d_{\pi_e}(s,a)}{d_{\mathcal{D}}(s,a)}\right)$

where equality holds only if ϕ is the identity function i.e. $\phi(s) = s, \forall s \in \mathcal{S}$ and/or if $\forall s_1, s_2 \in \mathcal{S}$ such that $\forall s_1, s_2 \in s^\phi$ and for a given action a, $\frac{d_{\pi_e}(s_1,a)}{d_{\mathcal{D}}(s_1,a)} = \frac{d_{\pi_e}(s_2,a)}{d_{\mathcal{D}}(s_2,a)}, \forall s^\phi \in \mathcal{S}^\phi, a \in \mathcal{A}$. Through this example, we have shown how projecting $\mathcal{S} \to \mathcal{S}^\phi$ can lower the variance of density ratios.

3.2 MIS OPE with True Abstract Ratios

We now present the statistical properties of our estimator assuming it has access to the true abstract state-action ratios. Due to space constraints, we defer proofs to Appendix A.3.

We prove our abstract estimator is unbiased (Theorem 4 in Appendix A.3) and strongly consistent (Theorem 2 and Corollary 1):

Theorem 2. *Our estimator, $\hat{\rho}(\pi_e^\phi)$, given in Equation 3 is an asymptotically consistent estimator of $\rho(\pi_e)$ in terms of MSE:* $\lim_{N\to\infty} \mathbb{E}[(\hat{\rho}(\pi_e^\phi) - \rho(\pi_e))^2] = 0$.

We also compare the variances between our abstract estimator and the ground equivalent. If we assume that $\mathcal{D}$ is i.i.d as done in previous work Sutton et al. (2008); Uehara et al. (2019); Nachum et al. (2019); Zhang et al. (2020a), then $\text{Var}(\hat{\rho}(\pi_e^\phi)) \leq \text{Var}(\hat{\rho}(\pi_e))$, where equality holds only under the same conditions as described above for Theorem 1. In general, however, variance reduction depends on the covariances between the weighted per-step rewards Liu et al. (2019):

Theorem 3. *If Assumption 1 and if for any fixed* $1 \leq t < k \leq T$, $Cov\left(\frac{d_{\pi_e^\phi}(s_t^\phi,a_t)}{d_{\mathcal{D}^\phi}(s_t^\phi,a_t)} r^\phi(s_t^\phi,a_t), \frac{d_{\pi_e^\phi}(s_k^\phi,a_k)}{d_{\mathcal{D}^\phi}(s_k^\phi,a_k)} r^\phi(s_k^\phi,a_k)\right) \leq Cov\left(\frac{d_{\pi_e}(s_t,a_t)}{d_{\mathcal{D}}(s_t,a_t)} r(s_t,a_t), \frac{d_{\pi_e}(s_k,a_k)}{d_{\mathcal{D}}(s_k,a_k)} r(s_k,a_k)\right)$ *hold then* $Var(\hat{\rho}(\pi_e^\phi)) \leq Var(\hat{\rho}(\pi_e))$.

3.3 Estimating the Abstract Ratios

Thus far, we have assumed access to the true abstract ratios. However, in practice, these ratios are unknown and must be estimated from $\mathcal{D}^\phi$. In this section, we highlight the challenges in estimating the abstract ratios and identify sufficient conditions on ϕ that allow for accurate estimation. Once ϕ satisfies these conditions, any off-the-shelf state-action density estimation method can be used to estimate the abstract ratios. In this work, we focus on showing that DICE estimates the true abstract ratios. We note that the following new conditions on ϕ are only needed for unbiased estimation of the abstract ratios; accurate OPE on the abstracted MDP can still be realized under only Assumption 1.

We first observe that evaluating π_e using $\mathcal{D}$ is equivalent to evaluating π_e^ϕ using $\mathcal{D}^\phi$ if $\mathcal{D}^\phi$ is generated from an abstract MDP with transition dynamics constructed according to w_{π_e}. This equivalence is given by the following proposition:

Proposition 1. *If Assumption 1 holds, the average reward of ground policy π executed in ground MDP $\mathcal{M}$, $\rho(\pi)$, is equal to the average reward of abstract policy π^ϕ executed in abstract MDP $\mathcal{M}^\phi$ constructed with w_π, $\rho(\pi^\phi)$. That is, $\rho(\pi) = \rho(\pi^\phi)$.*

Proposition 1 suggests that we can estimate $\rho(\pi_e)$ by applying any OPE algorithm to evaluate π_e^ϕ using $\mathcal{D}^\phi$. Unfortunately, there are two challenges which prevent us from doing so in the general case. Fortunately, there are special cases where unbiased estimation of the abstract ratios is still possible using only existing MIS algorithms.

Challenge 1: Transition dynamics distribution shift. In practice, Proposition 1 cannot be immediately applied since w_{π_e} is unknown. Additionally, our off-policy data $\mathcal{D}^\phi := \{(s^\phi, a, r^\phi, s'^\phi)\}$ is distributed in the following way: $(s^\phi, a) \sim d_{\mathcal{D}^\phi}, r^\phi \sim \mathcal{R}^\phi, s'^\phi \sim P^\phi_{w_\mathcal{D}}(s^\phi, a)$ where $P^\phi_{w_\mathcal{D}}$ are the transition dynamics of the abstract MDP constructed with $w_\mathcal{D}$ as the state-weighting function. Thus, in addition to the original policy distribution shift problem, we also encounter a *transition dynamics distribution shift* problem due to the projection where we want to evaluate π_e^ϕ in an MDP with $P^\phi_{w_{\pi_e}}$, but we only have samples of data generated in an MDP with $P^\phi_{w_\mathcal{D}}$. Since the transition dynamics are unknown, we cannot correct its distribution shift as we would correct for policy distribution shift. However, one condition on ϕ that will avoid the transition distribution shift is:

Assumption 2 (Transition dynamics similarity)**.** $\forall s_1, s_2 \in \mathcal{S}$ *such that* $s_1, s_2 \in s^\phi$, $\forall a \in \mathcal{A}, x \in \mathcal{S}^\phi$, *we have* $\sum_{s' \in \phi^{-1}(x)} P(s'|s_1, a) = \sum_{s' \in \phi^{-1}(x)} P(s'|s_2, a)$.

Together with Assumption 1, ϕ is now the so-called *bisimulation* abstraction Ferns et al. (2011); Castro (2020). This property of ϕ eliminates the transition dynamics distribution shift since now $P^\phi_{w_{\pi_e}}(s'^\phi|s^\phi, a) = P^\phi_{w_\mathcal{D}}(s'^\phi|s^\phi, a) = P^\phi(s'^\phi|s^\phi, a) = \sum_{s' \in s'^\phi} P(s'|s, a) \forall s \in s^\phi, \forall a \in \mathcal{A}$ (applying Assumption 2 and definition of P^ϕ from Section 2.3). Thus, we have $\mathcal{D}^\phi$ distributed with $P^\phi_{w_{\pi_e}} = P^\phi$, which then allows us to apply any MIS algorithm to compute the abstract state-action ratios using $\mathcal{D}^\phi$.

Challenge 2: Inaccessible π_e^ϕ. To the best of our knowledge, all existing MIS algorithms require access to the evaluation policy to estimate the density ratios. However, in the abstract MDP, π_e^ϕ is inaccessible since w_{π_e} is unknown. To overcome this issue, we identify the following condition on ϕ:

Assumption 3 (π_e action-distribution equality)**.** $\forall s_1, s_2 \in \mathcal{S}$ *such that* $s_1, s_2 \in s^\phi$, $\pi_e(s_1) = \pi_e(s_2)$

Assumption 3 then gives us $\pi_e^\phi(s^\phi) = \pi_e(s), \forall s \in s^\phi$ (applying Assumption 3 and definition of π^ϕ from Section 2.3), which allows us to simulate sampling from $\pi_e^\phi(\cdot|\phi(s))$ by just sampling from $\pi_e(\cdot|s)$.

Given a ϕ with these properties, we can compute the abstract ratios needed to estimate $\rho(\pi_e)$ by applying a suitable MIS algorithm to $\mathcal{D}^\phi$. We use BestDICE Yang et al. (2020), and call our algorithm AbstractBestDICE, which solves the following optimization problem:

$$\min_{\nu,\lambda} \max_{\zeta} J(\nu, \zeta, \lambda) := -\mathbb{E}_{\mathcal{D}^\phi}\left[\frac{1}{2}\zeta(s^\phi, a)^2\right] + \mathbb{E}_{\mathcal{D}^\phi}\left[\zeta(s^\phi, a)\left(\gamma\, \mathbb{E}_{a' \sim \pi_e^\phi}[\nu(s'^\phi, a')] - \nu(s^\phi, a) - \lambda\right)\right] + (1-\gamma)\, \mathbb{E}_{s_0^\phi \sim d_{0^\phi}, a_0 \sim \pi_e^\phi}[\nu(s_0^\phi, a_0)] + \lambda \tag{4}$$

where, $\nu : \mathcal{S}^\phi \times \mathcal{A} \to \mathbb{R}$, $\lambda \in \mathbb{R}$, and $\zeta : \mathcal{S}^\phi \times \mathcal{A} \to \mathbb{R}_{\geq 0}$. The solution to the optimization, $\zeta^*(s^\phi, a) = d_{\pi_e^\phi}(s^\phi, a)/d_{\mathcal{D}^\phi}(s^\phi, a)$ is the true abstract ratios. Since the derivation of AbstractBestDICE follows the same steps as BestDICE, we defer it to Appendix A.4.

We note that it may be difficult to validate Assumptions 2 and 3 in practice, which may result in loss of the consistency guarantee of Theorem 2. Nevertheless, in Section 4 we show that AbstractBestDICE leads to accurate OPE in high-dimensional state-spaces even when assumptions may not hold.

4 Empirical Study

We will now show how projecting $\mathcal{S} \to \mathcal{S}^\phi$ can produce more accurate OPE estimates in practice. We answer the following questions:

1. Do the true abstract ratios produce lower variance OPE estimates than the true ground ratios?

2. Does AbstractBestDICE: (a) compute the true ratios when Assumptions 1, 2, and 3 are satisfied and (b) produce data-efficient and stable estimates of $\rho(\pi_e)$ even when Assumptions 2 and 3 fail to hold?

4.1 Empirical Setup

In this section, we describe the algorithms and domains of our empirical study. Due to space constraints, we defer details regarding policies, hyperparamer tuning, etc. to the appendix (A.5 and A.6).

4.1.1 Algorithms

We compare AbstractBestDICE to ground BestDICE Yang et al. (2020). As also reported by Yang et al. (2020); Fu et al. (2021), we found in preliminary experiments that BestDICE performed much better than other variants of DICE such as DualDICE Nachum et al. (2019), GradientDICE Zhang et al. (2020b), etc.

4.1.2 Domains

We focus on high-dimensional state-space tasks, which have been known to be particularly challenging for DICE methods Fu et al. (2021). For each environment below we specify a fixed ϕ. For Reacher, Walker2D, Pusher, and AntUMaze, ϕ satisfies only Assumption 1.

- **TwoPath MDP**. This MDP is pictured in Figure 2. In this domain, Assumptions 1, 2, and 3 are satisfied. We also run the same experiments for when these assumptions are violated (see Appendix A.5). All trajectories are 100 time-steps long.
- **Reacher** Brockman et al. (2016). A robotic arm tries to move to a goal location. Here, $s \in \mathbb{R}^{11}$ and $a \in \mathbb{R}^2$. Since the reward function is the Euclidean distance between the arm and goal, ϕ extracts only the arm-to-goal vector, and angular velocities from the ground state, resulting in $s^\phi \in \mathbb{R}^4$. All trajectories are 200 time-steps long.
- **Walker2D** Brockman et al. (2016). A bi-pedal robot tries to move as fast as possible. Here, $s \in \mathbb{R}^{18}$ and $a \in \mathbb{R}^6$. We use the Euclidean distance from the start location as the reward function and use a ϕ that extracts x and z coordinates and top angle of the walker's body, resulting in $s^\phi \in \mathbb{R}^3$. All trajectories are 500 time-steps long.
- **Pusher** Brockman et al. (2016). A robotic arm tries to push an object to a goal location. Here, $s \in \mathbb{R}^{23}$ and $a \in \mathbb{R}^7$. Since the reward function is the Euclidean distance between object and arm and object and goal, ϕ extracts only object-to-arm and object-to-goal vectors, resulting in $s^\phi \in \mathbb{R}^6$. All trajectories are 300 time-steps long.
- **AntUMaze** Fu et al. (2020). This sparse-reward task requires an ant to move from one end of the U-shaped maze to the other end. Here, $s \in \mathbb{R}^{29}$ and $a \in \mathbb{R}^8$. We use the "play" version where the goal location is fixed. Since the reward function is $+1$ only if the 2D location of the ant is at a certain Euclidean distance from the 2D goal location, ϕ extracts only the 2D coordinates of the ant, resulting in $s^\phi \in \mathbb{R}^2$. All trajectories are 500 time-steps long.

4.2 Empirical Results

In this section, we describe our main empirical results; additional experiments can be found in appendix A.6.

4.2.1 True Ratios for OPE

We conduct an experiment on the TwoPath MDP to estimate $\rho(\pi_e)$ where we apply the ground estimator given in Equation (1) and our abstract estimator given in Equation (3), assuming *both have access to their respective true ratios*. The results of this experiment are illustrated in Figure 3(a). We can observe that the abstract estimator with the true abstract ratios produces substantially more data-efficient and lower variance OPE estimates for different batch sizes compared to the ground equivalent.

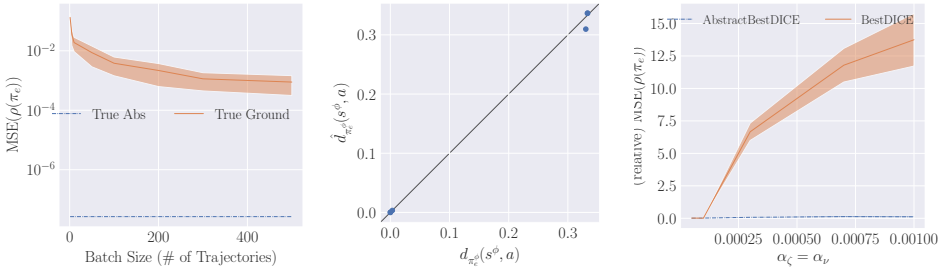
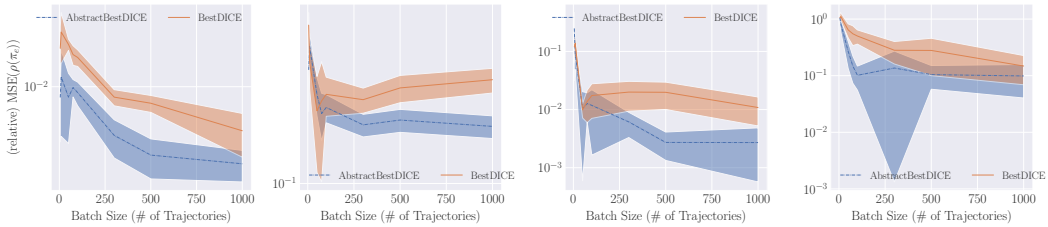


Figure 3: True ratio experiments. Figure 3(a): MSE vs. Batch size (# of trajectories). The vertical axis axis is log-scaled. Errors are computed over 15 trials with 95% confidence intervals. Lower is better. Since $\rho(\pi_e) = \rho(\pi_b)$ due to their symmetry, we use regular MSE instead of relative. Figure 3(b): Estimated $\hat{d}_{\pi_e^\phi}(s^\phi, a)$ (vertical axis) vs. true $d_{\pi_e^\phi}$ (horizontal axis). Values are averaged over 15 trials. We expect the dots to be as close to the black line as possible. Each dot is for each (s^ϕ, a). Dots are only at extreme ends due to choice of π_e, $\pi_\mathcal{D}$, and ToyMDP design. Only 4 out of 6 dots are visible due to overlap between the dots for (s_1^ϕ, a_1) and (s_2^ϕ, a_1), and (s_1^ϕ, a_0) and (s_2^ϕ, a_0). Figure 3(c): Robustness of BestDICE and AbstractBestDICE to hyperparameters on the Pusher domain for batch size (# of trajectories) of 50. Errors are computed over 15 trials with 95% confidence intervals. Lower is better.

4.2.2 True Ratio Estimation

To verify if AbstractBestDICE accurately estimates the true ratios we conduct the following experiment on the TwoPath MDP. We give AbstractBestDICE data of batch size 300 to estimate the abstract ratios $\hat{\zeta}^\phi$ and then use $\hat{\zeta}^\phi$ to estimate the abstract state-action densities of π_e^ϕ, $\hat{d}_{\pi_e^\phi}(s^\phi, a) = \hat{\zeta}^\phi(s^\phi, a) d_{\pi_\mathcal{D}^\phi}(s^\phi, a)$, where we have access to the true $d_{\pi_\mathcal{D}^\phi}(s^\phi, a)$. We then compare $\hat{d}_{\pi_e^\phi}$ to the true $d_{\pi_e^\phi}$, which we compute using a batch size of 300 trajectories collected from π_e^ϕ roll-outs, using a correlation plot shown in Figure 3(b). From the figure we can see AbstractBestDICE accurately estimates the abstract state-action density ratios. When assumptions are violated, however, ratio estimation accuracy can reduce (see Appendix A.5).



(a) Reacher (b) Walker2D (c) Pusher (d) AntUMaze

Figure 4: Relative MSE vs. Batch Size (# of trajectories). Vertical axis is log-scaled. Errors are computed over 15 trials with 95% confidence intervals. Lower is better.

4.2.3 Data-Efficiency

Figure 4 shows the results of our (relative) MSE vs. batch size experiment for the function approximation case. For a given batch size, we train each algorithm for 100k epochs with different hyperparameters sets, record the (relative) MSE on the last epoch by each hyperparameter set, and plot the lowest MSE achieved by these hyperparameter sets. We find that AbstractBestDICE is able to achieve lower MSE than BestDICE for a given batch size. We note that while hyperparameter tuning is difficult in OPE, in this experiment, we aim to evaluate the performance of each algorithm assuming both had favorable hyperparameters.

4.2.4 Hyperparameter Robustness

Finally, we study the robustness of these algorithms to hyperparameters tuning. In practical OPE, hyperparameter tuning with respect to MSE is impractical since the true $\rho(\pi_e)$ is unknown Fu et al. (2021); Paine et al. (2020). Thus, we want OPE algorithms to be as robust as possible to hyperparameter tuning. The main hyperparameters for DICE are the learning rates of ζ and ν, α_ζ and α_ν. For these experiments, we focus on very small batch sizes, where we would expect high sensitivity. The results of this study are in Figure 3(c). We find that our algorithm has a less volatile MSE than BestDICE (also see appendix A.6 for more similar results). In a related experiment, we also find AbstractBestDICE can be more stable than BestDICE during training (see appendix A.6).

5 Related Work

MIS and Off-Policy Evaluation. There have been broadly three families of MIS algorithms in the OPE literature to estimate state-action density ratios. One is the DICE family, which includes: minimax-weight learning Uehara et al. (2019), DualDice Nachum et al. (2019), GenDICE Zhang et al. (2020a), GradientDICE Zhang et al. (2020b), and BestDICE Yang et al. (2020). In our work, we adapt BestDICE to estimate the abstract ratios. The second family of MIS algorithms is the COP-TD algorithm Hallak and Mannor (2017); Gelada and Bellemare (2019), which learns the state density ratios with an online TD-styled update. The third family is the variational power method Wen et al. (2020) algorithm which generalizes the power iteration method to estimate density ratios. While our focus has been on MIS algorithms, there are many other OPE algorithms such as model-based methods Zhang et al. (2021b); Hanna et al. (2017); Liu et al. (2018b), fitted-Q evaluation Le et al. (2019), doubly-robust methods Jiang and Li (2015); Thomas and Brunskill (2016), and IS Precup et al. (2000); Thomas (2015); Hanna et al. (2019); Thomas et al. (2015).

State Abstraction and Representation Learning. The literature on state abstraction is extensive Singh et al. (1994); Dietterich (1999); Ferns et al. (2011); Li et al. (2006); Abel (2020). However, much of this work has been exclusively focused on building a theory of abstraction and on learning optimal policies. A related topic to state abstraction is representation learning. Recently, there has been much work showing the importance of good representations for offline RL Wang et al. (2021); Yin et al. (2022); Geng et al. (2022); Zhan et al. (2022); Chen and Jiang (2022). To the best of our knowledge, no work has leveraged state abstraction techniques to improve the accuracy of OPE algorithms.

6 Summary and Future Work

In this work, we showed that we can improve the accuracy of OPE estimates by projecting the original ground state-space into a lower-dimensional abstract state-space using state abstraction and performing OPE in the resulting abstract Markov decision process. Our theoretical results proved that: 1) abstract state-action ratios have variance at most that of the ground ratios; and 2) the abstraction MIS OPE estimator is unbiased, strongly consistent, and can have lower variance than the ground equivalent. We then highlighted the challenges that arise when estimating the abstract ratios from data, identified sufficient conditions to overcome these issues, and adapted BestDICE into AbstractBestDICE to estimate the abstract ratios. In our empirical results, we obtained more accurate estimates with added hyperparameter robustness on difficult, high-dimensional state-space tasks.

There are several directions for future work. First, Assumptions 2 and 3 are strict. Further investigation is needed to see if these assumptions can be relaxed. Second, we assumed the abstraction function was given. It would be interesting to leverage existing ideas Gelada et al. (2019); Zhang et al. (2021a) to learn ϕ. Finally, we want to emphasize that this work instantiates the general abstraction + OPE direction. While this work focused exclusively on MIS algorithms, a promising direction will be to apply abstraction techniques to model-based, trajectory IS, and value-function based OPE.

Acknowledgements

The authors thank Nicholas Corrado, Ishan Durugkar, and Subhojyoti Mukherjee for reviewing earlier drafts of this work.

References

David Abel. *A Theory of Abstraction in Reinforcement Learning*. PhD thesis, Brown University, 2020.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *CoRR*, abs/1606.01540, 2016. URL http://arxiv.org/abs/1606.01540.

Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(06):10069–10076, Apr. 2020. doi: 10.1609/aaai.v34i06.6564. URL https://ojs.aaai.org/index.php/AAAI/article/view/6564.

Jinglin Chen and Nan Jiang. Offline reinforcement learning under value and density-ratio realizability: The power of gaps. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. URL https://openreview.net/forum?id=B5LzMd8jcg9.

Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual kernel embeddings. *CoRR*, abs/1607.04579, 2016. URL http://arxiv.org/abs/1607.04579.

Thomas G. Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition, 1999. URL https://arxiv.org/abs/cs/9905014.

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011. doi: 10.1137/10080484X. URL https://doi.org/10.1137/10080484X.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R. Zhang, Yutian Chen, Aviral Kumar, Cosmin Paduraru, Sergey Levine, and Thomas Paine. Benchmarks for deep off-policy evaluation. In *ICLR*, 2021. URL https://openreview.net/forum?id=kWSeGEeHvF8.

Carles Gelada and Marc G. Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3647–3655, Jul. 2019. doi: 10.1609/aaai.v33i01.33013647. URL https://ojs.aaai.org/index.php/AAAI/article/view/4246.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. Deepmdp: Learning continuous latent space models for representation learning. *CoRR*, abs/1906.02736, 2019. URL http://arxiv.org/abs/1906.02736.

Xinyang Geng, Kevin Li, Abhishek Gupta, Aviral Kumar, and Sergey Levine. Effective offline RL needs going beyond pessimism: Representations and distributional shift. In *Decision Awareness in Reinforcement Learning Workshop at ICML 2022*, 2022. URL https://openreview.net/forum?id=zEsshhtdwG4.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation, 2017. URL https://arxiv.org/abs/1702.07121.

Josiah Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2017.

Josiah Hanna, Scott Niekum, and Peter Stone. Importance sampling policy evaluation with an estimated behavior policy. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, June 2019.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning, 2015. URL https://arxiv.org/abs/1511.03722.

Hoang M. Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints, 2019. URL https://arxiv.org/abs/1903.08738.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for mdps. In *In Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics*, pages 531–539, 2006.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *CoRR*, abs/1810.12429, 2018a. URL http://arxiv.org/abs/1810.12429.

Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo Faisal, Finale Doshi-Velez, and Emma Brunskill. Representation balancing mdps for off-policy policy evaluation, 2018b. URL https://arxiv.org/abs/1805.09044.

Yao Liu, Pierre-Luc Bacon, and Emma Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling, 2019. URL https://arxiv.org/abs/1910.06508.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/cf9a242b70f45317ffd281241fa66502-Paper.pdf.

Tom Le Paine, Cosmin Paduraru, Andrea Michi, Çaglar Gülçehre, Konrad Zolna, Alexander Novikov, Ziyu Wang, and Nando de Freitas. Hyperparameter selection for offline reinforcement learning. *CoRR*, abs/2007.09055, 2020. URL https://arxiv.org/abs/2007.09055.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 759–766, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Satinder Singh, Tommi Jaakkola, and Michael Jordan. Reinforcement learning with soft state aggregation. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994. URL https://proceedings.neurips.cc/paper/1994/file/287e03db1d99e0ec2edb90d079e142f3-Paper.pdf.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Richard S. Sutton, Csaba Szepesvári, Alborz Geramifard, and Michael Bowling. Dyna-style planning with linear function approximation and prioritized sweeping. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, page 528–536, Arlington, Virginia, USA, 2008. AUAI Press. ISBN 0974903949.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015. doi: 10.1609/aaai.v29i1.9541. URL https://ojs.aaai.org/index.php/AAAI/article/view/9541.

Philip S. Thomas. *Safe Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 2015.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning, 2016.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation, 2019. URL https://arxiv.org/abs/1910.12809.

Cameron Voloshin, Hoang Minh Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=IsK8iKbL-I.

Ruosong Wang, Yifan Wu, Ruslan Salakhutdinov, and Sham M. Kakade. Instabilities of offline rl with pre-trained neural representation, 2021. URL https://arxiv.org/abs/2103.04947.

Junfeng Wen, Bo Dai, Lihong Li, and Dale Schuurmans. Batch stationary distribution estimation. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/4ffb0d2ba92f664c2281970110a2e071-Paper.pdf.

Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6551–6561. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/488e4104520c6aab692863cc1dba45af-Paper.pdf.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism, 2022. URL https://arxiv.org/abs/2203.05804.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2730–2775. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/zhan22a.html.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=-2FCwDKRREu.

Michael R Zhang, Thomas Paine, Ofir Nachum, Cosmin Paduraru, George Tucker, ziyu wang, and Mohammad Norouzi. Autoregressive dynamics models for offline policy evaluation and optimization. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=kmqjgSNXby.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values, 2020a. URL https://arxiv.org/abs/2002.09072.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. GradientDICE: Rethinking generalized offline estimation of stationary values. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11194–11203. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/zhang20r.html.

A Appendix

A.1 Preliminaries

This section provides the supporting lemmas and definitions that we leverage to prove our lemmas and theorems.

Definition 2 (Almost Sure Convergence)**.** *A sequence of random variables, $(X_n)_{n=1}^{\infty}$, almost surely converges to the random variable, X if*

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$$

We write $X_n \overset{a.s.}{\to} X$ to denote that the sequence $(X_n)_{n=1}^{\infty}$ converges almost surely to X.

Definition 3 ((Strongly) Consistent Estimator)**.** *Let θ be a real number and $(\hat{\theta}_n)_{n=1}^{\infty}$ be an infinite sequence of random variables. We call $\hat{\theta}_n$ a (strongly) consistent estimator of θ if and only if $\hat{\theta}_n \overset{a.s.}{\to} \theta$.*

Lemma 1. *If $(X_i)_{i=1}^{\infty}$ is a sequence of uniformly bounded real-valued random variables, then $X_n \overset{a.s.}{\to} X$ if and only if $\lim_{N\to\infty} \mathbb{E}[(X_n - X)^2] = 0$.*

Proof See Lemma 3 in Thomas and Brunskill (2016). □

A.2 Assumptions and Definitions

In the main paper, we provided the major assumptions required for our theoretical and empirical work relevant to abstraction and OPE. Here we provide supporting assumptions typically used in the OPE literature used for the theoretical analysis.

Assumption 4 (Coverage)**.** *For all $(s,a) \in \mathcal{S} \times \mathcal{A}$, if $\pi_e(a|s) > 0$ then $\pi_b(a|s) > 0$.*

Assumption 5 (Non-negative reward)**.** *We assume that the reward function is bounded between $[0, \infty)$.*

Definition 4 (Ground state normalized weightings)**.** *For a given policy π, each ground state $s \in s^{\phi}$, has a state aggregation weight, $w_{\pi}(s) = \frac{d_{\pi}(s)}{\sum_{s' \in \phi^{-1}(s^{\phi})} d_{\pi}(s')}$, where $d_{\pi}(s)$ is the discounted state-occupancy measure of π.*

A.3 Proofs

In the proofs below, we denote the collection of behavior policies that generated $\mathcal{D}$ with $\pi_{\mathcal{D}}$. That is, $\pi_{\mathcal{D}}$ is the conditional probability of an action occurring in a given state in the data. Similarly, we also have $\pi_{\mathcal{D}}^{\phi}$. These minor changes give us $d_{\mathcal{D}} = d_{\pi_{\mathcal{D}}}$ and $d_{\mathcal{D}^{\phi}} = d_{\pi_{\mathcal{D}}^{\phi}}$.

Lemma 2. *For an arbitrary function, f, $\mathbb{E}_{s^{\phi}\sim d_{\pi^{\phi}}, a\sim\pi^{\phi}}\left[f(s^{\phi}, a)\right] = \mathbb{E}_{s\sim d_{\pi}, a\sim\pi}\left[f(\phi(s), a)\right]$.*

Proof

$$\begin{aligned}
\mathbb{E}_{s^{\phi}\sim d_{\pi^{\phi}}, a\sim\pi^{\phi}}\left[f(s^{\phi}, a)\right] &= \sum_{s^{\phi}, a} d_{\pi^{\phi}}(s^{\phi})\pi^{\phi}(a|s^{\phi}) f(s^{\phi}, a) \\
&\overset{\text{(a)}}{=} \sum_{s^{\phi}, a} d_{\pi^{\phi}}(s^{\phi}) \sum_{s\in\phi^{-1}(s^{\phi})} \frac{\pi(a|s) d_{\pi}(s)}{d_{\pi^{\phi}}(s^{\phi})} f(s^{\phi}, a) \\
&= \sum_{s^{\phi}, a} \sum_{s\in\phi^{-1}(s^{\phi})} \pi(a|s) d_{\pi}(s) f(s^{\phi}, a) \\
&= \sum_{s, a} \pi(a|s) d_{\pi}(s) f(\phi(s), a) \\
\mathbb{E}_{s^{\phi}\sim d_{\pi}^{\phi}, a\sim\pi^{\phi}}\left[f(s^{\phi}, a)\right] &= \mathbb{E}_{s\sim d_{\pi}, a\sim\pi}\left[f(\phi(s), a)\right]
\end{aligned}$$

where (a) is due to Definition 4 and we can replace s^{ϕ} with $\phi(s)$ when we know $s \in s^{\phi}$. □

Theorem 1. *Var* $\left(\frac{d_{\pi_e^\phi}(s^\phi,a)}{d_{\mathcal{D}^\phi}(s^\phi,a)}\right) \leq$ *Var* $\left(\frac{d_{\pi_e}(s,a)}{d_{\mathcal{D}}(s,a)}\right)$

Proof

Before comparing the variances, we note that due to Assumption 4 and Lemma 2:

$$\mathbb{E}_{s\sim d_{\pi_\mathcal{D}},a\sim\pi_\mathcal{D}}\left[\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right] = \mathbb{E}_{s\sim d_{\pi_\mathcal{D}^\phi},a\sim\pi_\mathcal{D}^\phi}\left[\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}\right] = 1$$

Denote, $V^g :=$ Var $\left(\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right)$ and $V^\phi :=$ Var $\left(\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}\right)$. Now consider the difference between the two variances.

$$\begin{aligned}
D &= V^g - V^\phi \\
&= \text{Var}\left(\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right) - \text{Var}\left(\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}\right) \\
&= \mathbb{E}_{s\sim d_{\pi_\mathcal{D}},a\sim\pi_\mathcal{D}}\left[\left(\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right)^2\right] - \mathbb{E}_{s^\phi\sim d_{\pi_\mathcal{D}^\phi},a\sim\pi_\mathcal{D}^\phi}\left[\left(\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}\right)^2\right] \\
&= \sum_{s,a} d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)\left(\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right)^2 - \sum_{s^\phi,a} d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)\left(\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}\right)^2 \\
&= \sum_{s^\phi,a}\left(\sum_{s\in s^\phi} d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)\left(\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right)^2 - d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)\left(\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}\right)^2\right)
\end{aligned}$$

We can analyze this difference by looking at one abstract state and one action and all the states that belong to it. That is, for a fixed abstract state, s^ϕ, and fixed action, a, we have:

$$\begin{aligned}
D' &= \sum_{s\in s^\phi} d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)\left(\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right)^2 - \left(d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)\left(\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}\right)^2\right) \\
&= \left(\sum_{s\in s^\phi}\frac{(d_{\pi_e}(s)\pi_e(a|s))^2}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right) - \left(\frac{(d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi))^2}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}\right) \\
&\overset{\text{(a)}}{=} \left(\left(\sum_{s\in s^\phi}\frac{(d_{\pi_e}(s)\pi_e(a|s))^2}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right) - \left(\frac{(d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)^2}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}\right)\right)
\end{aligned}$$

where (a) is due to Definition 4.

If we can show that $D' \geq 0$ for all possible sizes of $|s^\phi|$, we will the have the original difference, D, is a sum of only non-negative terms, thus proving Theorem 1. We will prove $D' \geq 0$ by inductive proof on the size of $|s^\phi|$ from 1 to some $n \leq |\mathcal{S}|$.

Let our statement to prove, $P(n)$ be that $D' \geq 0$ where $n = |s^\phi|$. This is trivially true for $P(1)$ where the ground state equals the abstract state. Now consider the inductive hypothesis, $P(n)$ is true for $n \geq 1$. Now with the inductive step, we must show that $P(n+1)$ is true given $P(n)$ is true. Starting with the inductive hypothesis:

$$D'' = \underbrace{\left(\sum_{s \in s^\phi} \frac{(d_{\pi_e}(s)\pi_e(a|s))^2}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right)}_{S} - \left(\frac{\overbrace{((d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi))^2}^{C}}{\underbrace{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}_{C'}}\right) \geq 0$$

We define $S := \left(\sum_{s \in s^\phi} \frac{(d_{\pi_e}(s)\pi_e(a|s))^2}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right)$, $C := (d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)$, and $C' := d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)$. After making the substitutions, we have:

$$C^2 \leq SC' \tag{5}$$

We have the above result holding true for when the $|s^\phi| = n$. Now consider the inductive step in relation to the inductive hypothesis where a new state, s_{n+1} is added to the abstract state. We have the following difference:

$$D'' = S + \frac{(d_{\pi_e}(s_{n+1})\pi_e(a|s_{n+1}))^2}{d_{\pi_\mathcal{D}}(s_{n+1})\pi_\mathcal{D}(a|s_{n+1})} - \frac{(C + d_{\pi_e}(s_{n+1})\pi_e(a|s_{n+1}))^2}{C' + d_{\pi_\mathcal{D}}(s_{n+1})\pi_\mathcal{D}(a|s_{n+1})}$$

For ease in notation, let $x = d_{\pi_e}(s_{n+1})\pi_e(a|s_{n+1})$ and $y = d_{\pi_\mathcal{D}}(s_{n+1})\pi_\mathcal{D}(a|s_{n+1})$. The above difference is then:

$$\begin{aligned}
D'' &= S + \frac{x^2}{y} - \frac{(C+x)^2}{C'+y} \\
&= \frac{Sy + x^2}{y} - \frac{(C+x)^2}{C'+y} \\
&= \frac{1}{y(C'+y)}((Sy + x^2)(C'+y) - (C+x)^2 y) \\
&= \frac{1}{y(C'+y)}(SyC' + Sy^2 + x^2C' + x^2y - C^2y - x^2y - 2Cxy) \\
&= \frac{1}{y(C'+y)}(SyC' + Sy^2 + x^2C' - C^2y - 2Cxy)
\end{aligned}$$

The above difference, D'', is minimized most when C is as large as possible. From the inductive hypothesis, we have $C \leq \sqrt{SC'}$. The minimum difference can be written as:

$$\begin{aligned}
D'' &= \frac{1}{y(C'+y)}(Sy^2 + x^2C' - 2\sqrt{SC'}xy) \\
&= \frac{1}{y(C'+y)}(y\sqrt{S} - x\sqrt{C'})^2 \\
&\geq 0
\end{aligned}$$

So we have $D'' \geq 0$ for $|s^\phi| = n+1$, which means $D' \geq 0$. We have showed that $P(n)$ is true for all n. We now have the original difference, D, to be a sum of non-negative terms after performing this same grouping for all abstract states and actions, which results in:

$$\text{Var}\left(\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}\right) \leq \text{Var}\left(\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right)$$

Thus, we have:

$$\text{Var}\left(\frac{d_{\pi_e^\phi}(s^\phi, a)}{d_{\pi_\mathcal{D}^\phi}(s^\phi, a)}\right) \leq \text{Var}\left(\frac{d_{\pi_e}(s, a)}{d_{\pi_\mathcal{D}}(s, a)}\right)$$

□

Proposition 1. *If Assumption 1 holds, the average reward of ground policy π executed in ground MDP $\mathcal{M}$, $\rho(\pi)$, is equal to the average reward of abstract policy π^ϕ executed in abstract MDP $\mathcal{M}^\phi$ constructed with w_π, $\rho(\pi^\phi)$. That is, $\rho(\pi) = \rho(\pi^\phi)$.*

Proof Consider the definition of R^ϕ_π:

$$
\begin{aligned}
\rho(\pi^\phi) &= \sum_{s^\phi,a} d_{\pi^\phi}(s^\phi)\pi^\phi(a|s^\phi)r(s^\phi,a) \\
&= \sum_{s^\phi,a}\left(\left(\sum_{s\in\phi^{-1}(s^\phi)} d_\pi(s)\right)\left(\sum_{s\in\phi^{-1}(s^\phi)} \pi(a|s)w_\pi(s)\right) r(s^\phi,a)\right) \\
&\overset{\text{(a)}}{=} \sum_{\phi(s),a}\left(\left(\sum_{s\in\phi^{-1}(s^\phi)} d_\pi(s)\right)\left(\sum_{s\in\phi^{-1}(s^\phi)} \frac{\pi(a|s)d_\pi(s)}{\sum_{s'\in\phi^{-1}(s^\phi)} d_\pi(s')}\right) r(s^\phi,a)\right) \\
&\overset{\text{(b)}}{=} \sum_{\phi(s),a}\left(\sum_{s\in\phi^{-1}(s^\phi)} \pi(a|s)d_\pi(s)\right) r(s^\phi,a) \\
&\overset{\text{(c)}}{=} \sum_{\phi(s),a}\left(\sum_{s\in\phi^{-1}(s^\phi)} \pi(a|s)d_\pi(s)r(s,a)\right) \\
&= \sum_{s,a} \pi(a|s)d_\pi(s)r(s,a) \\
\rho(\pi^\phi) &= \rho(\pi)
\end{aligned}
$$

where (a) is due to Definition 4, (b) is due to Definition 4 and Assumption 1 □

Theorem 4. *If Assumption 1 holds, our estimator, $\hat{\rho}(\pi_e^\phi)$ as defined in Equation 3, is an unbiased estimator of $\rho(\pi_e)$.*

Proof

We first consider the expectation of a single sample, $X = \frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)} r^\phi(s^\phi,a)$:

$$
\begin{aligned}
\mathbb{E}_{s\sim d_{\pi_\mathcal{D}},a\sim\pi_\mathcal{D}}[X] &= \sum_{s,a} d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)} r^\phi(s^\phi,a) \\
&\overset{(a)}{=} \sum_{s^\phi,a}\sum_{s\in\phi^{-1}(s^\phi)} d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)} r^\phi(s^\phi,a) \\
&\overset{(b)}{=} \sum_{s^\phi,a}\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}\sum_{s\in\phi^{-1}(s^\phi)} d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s) r^\phi(s^\phi,a) \\
&\overset{(c)}{=} \sum_{s^\phi,a}\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)} r^\phi(s^\phi,a)\sum_{s\in\phi^{-1}(s^\phi)} d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s) \\
&= \sum_{s^\phi,a}\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)} r^\phi(s^\phi,a)\sum_{s'\in\phi^{-1}(s^\phi)} d_{\pi_\mathcal{D}}(s')\sum_{s\in\phi^{-1}(s^\phi)}\frac{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}{\sum_{s'\in\phi^{-1}(s^\phi)} d_{\pi_\mathcal{D}}(s')} \\
&= \sum_{s^\phi,a}\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)} r^\phi(s^\phi,a)(d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)) \\
&= \sum_{s^\phi,a} d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi) r^\phi(s^\phi,a) \\
&\overset{(d)}{=} \rho(\pi_e^\phi) \\
\mathbb{E}_{s\sim d_{\pi_\mathcal{D}},a\sim\pi_\mathcal{D}}[X] &\overset{(e)}{=} \rho(\pi_e)
\end{aligned}
$$

where (c) is due to Definition 4 and Assumption 1, (d) is due to Assumption 4, and (e) is due to Proposition 1.

We have the bias defined as:

$$
\begin{aligned}
\text{Bias}[\hat{\rho}(\pi_e^\phi)] &= \mathbb{E}_{s\sim d_{\pi_\mathcal{D}},a\sim\pi_\mathcal{D}}[\hat{\rho}(\pi_e^\phi)] - R_{\pi_e} \\
&= \mathbb{E}_{s\sim d_{\pi_\mathcal{D}},a\sim\pi_\mathcal{D}}\left[\frac{1}{mT}\sum_{i=1}^{mT}\frac{d_{\pi_e^\phi}(s_i^\phi)\pi_e^\phi(a_i|s_i^\phi)}{d_{\pi_\mathcal{D}^\phi}(s_i^\phi)\pi_\mathcal{D}^\phi(a_i|s_i^\phi)} r^\phi(s_i^\phi,a_i)\right] - R_{\pi_e} \\
&\overset{(a)}{=} \frac{1}{mT}\sum_{i=1}^{mT}\mathbb{E}_{s_i\sim d_{\pi_\mathcal{D}},a_i\sim\pi_\mathcal{D}}\left[\frac{d_{\pi_e^\phi}(s_i^\phi)\pi_e^\phi(a_i|s_i^\phi)}{d_{\pi_\mathcal{D}^\phi}(s_i^\phi)\pi_\mathcal{D}^\phi(a_i|s_i^\phi)} r^\phi(s_i^\phi,a_i)\right] - R_{\pi_e} \\
&\overset{(b)}{=} \left(\frac{1}{mT}\sum_{i=1}^{mT} R_{\pi_e}\right) - R_{\pi_e} \\
&= \rho(\pi_e) - \rho(\pi_e) \\
\text{Bias}[\hat{\rho}(\pi_e^\phi)] &= 0
\end{aligned}
$$

where (a) is due to linearity of expectation and (b) is due to expectation of a single sample. □

Theorem 2. *Our estimator, $\hat{\rho}(\pi_e^\phi)$, given in Equation 3 is an asymptotically consistent estimator of $\rho(\pi_e)$ in terms of MSE:* $\lim_{N\to\infty}\mathbb{E}[(\hat{\rho}(\pi_e^\phi)-\rho(\pi_e))^2] = 0$.

Proof We have the MSE of $\hat{\rho}(\pi_e^\phi)$ w.r.t $\rho(\pi_e)$ defined in terms of the bias and variance as follows:

$$\begin{aligned}\text{MSE}(\hat{\rho}(\pi_e^\phi)) = \mathbb{E}[(\hat{\rho}(\pi_e^\phi) - \rho(\pi_e))^2] &= \text{Var}[\hat{\rho}(\pi_e^\phi)] + (\text{Bias}[\hat{\rho}(\pi_e^\phi)])^2 \\ &\overset{\text{(a)}}{=} \text{Var}[\hat{\rho}(\pi_e^\phi)] \\ &= \frac{1}{(mT)^2}\text{Var}\left(\sum_{i=1}^{mT} \frac{d_{\pi_e^\phi}(s_i^\phi)\pi_e^\phi(a_i|s_i^\phi)}{d_{\pi_\mathcal{D}^\phi}(s_i^\phi)\pi_\mathcal{D}^\phi(a_i|s_i^\phi)} r^\phi(s_i^\phi, a_i)\right)\end{aligned}$$

where (a) is because $\hat{\rho}$ is an unbiased estimator as shown in Theorem 4.

Due to Assumptions 4 and 5, $\left(\sum_{i=1}^{mT} \frac{d_{\pi_e^\phi}(s_i^\phi)\pi_e^\phi(a_i|s_i^\phi)}{d_{\pi_\mathcal{D}^\phi}(s_i^\phi)\pi_\mathcal{D}^\phi(a_i|s_i^\phi)} r^\phi(s_i^\phi, a_i)\right)$ is a bounded value. Thus, as $mT \to \infty$, $\text{Var}[\hat{\rho}(\pi_e^\phi)] \to 0$. We then have $\lim_{mT\to\infty} \mathbb{E}[(\hat{\rho}(\pi_e^\phi) - \rho(\pi_e))^2] = 0$. Thus, the estimator $\hat{\rho}(\pi_e^\phi)$ is consistent in MSE. □

Corollary 1. *If Assumption 1 holds, then our estimator, $\hat{\rho}(\pi_e^\phi)$ as defined in Equation 3 is an asymptotically strongly consistent estimator of $\rho(\pi_e)$.*

Proof Theorem 2 showed that $\hat{\rho}(\pi_e^\phi)$ is consistent in terms of MSE. Then by applying Lemma 1, we have $\hat{\rho}(\pi_e^\phi)$ to be an asymptotically strongly consistent estimator of $\rho(\pi_e)$. That is, $\hat{\rho}(\pi_e^\phi) \overset{a.s.}{\to} \rho(\pi_e)$. □

Note that this proof is very similar to the one given in Theorem 1, where the main the difference is that we are analyzing the variance of an OPE estimator on given batch of data.

Theorem 3. *If Assumption 1 and if for any fixed $1 \le t < k \le T$, $\text{Cov}\left(\frac{d_{\pi_e^\phi}(s_t^\phi,a_t)}{d_{\mathcal{D}^\phi}(s_t^\phi,a_t)} r^\phi(s_t^\phi, a_t), \frac{d_{\pi_e^\phi}(s_k^\phi,a_k)}{d_{\mathcal{D}^\phi}(s_k^\phi,a_k)} r^\phi(s_k^\phi, a_k)\right) \le \text{Cov}\left(\frac{d_{\pi_e}(s_t,a_t)}{d_\mathcal{D}(s_t,a_t)} r(s_t, a_t), \frac{d_{\pi_e}(s_k,a_k)}{d_\mathcal{D}(s_k,a_k)} r(s_k, a_k)\right)$ hold then $\text{Var}(\hat{\rho}(\pi_e^\phi)) \le \text{Var}(\hat{\rho}(\pi_e))$.*

Proof

We first consider the general form of the variance of the baseline estimator, $\hat{\rho}(\pi_e)$ (and the similar form applies to $\hat{\rho}(\pi_e^\phi)$). For simplicity in notation, we take the batch size, $m = 1$, but the analysis holds for general m.

$$\begin{aligned}\text{Var}[\hat{\rho}(\pi_e)] &= \text{Var}\left(\frac{1}{T}\sum_{t=1}^{T} \frac{d_{\pi_e}(s_t)\pi_e(a_t|s_t)}{d_{\pi_\mathcal{D}}(s_t)\pi_\mathcal{D}(a_t|s_t)} r(s_t, a_t)\right) \\ &= \frac{1}{T^2}\left(\sum_{t=1}^{T} \underbrace{\text{Var}\left(\frac{d_{\pi_e}(s_t)\pi_e(a_t|s_t)}{d_{\pi_\mathcal{D}}(s_t)\pi_\mathcal{D}(a_t|s_t)} r(s_t, a_t)\right)}_{V_t^g} + 2\sum_{t<k} \text{Cov}\left(\frac{d_{\pi_e}(s_t,a_t)}{d_{\pi_\mathcal{D}}(s_t,a_t)} r(s_t, a_t), \frac{d_{\pi_e}(s_k,a_k)}{d_{\pi_\mathcal{D}}(s_k,a_k)} r(s_k, a_k)\right)\right)\end{aligned}$$

Here we have V_t^g to be the variance of a single sample at time t. Similarly, for our estimator, we have $V_t^\phi = \text{Var}\left(\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)} r^\phi(s^\phi, a)\right)$. We will show that $V_t^\phi \le V_t^g$. We drop the subscript t for convenience since the analysis applies for all t.

Before comparing the variances, we note that due to Assumption 1 and 4, Lemma 2, and Proposition 1:

$$\mathbb{E}_{s\sim d_{\pi_\mathcal{D}}, a\sim\pi_\mathcal{D}}\left[\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)} r(s, a)\right] = \mathbb{E}_{s\sim d_{\pi_\mathcal{D}}, a\sim\pi_\mathcal{D}}\left[\frac{d_{\pi_e^\phi}(\phi(s))\pi_e^\phi(a|\phi(s))}{d_{\pi_\mathcal{D}^\phi}(\phi(s))\pi_\mathcal{D}^\phi(a|\phi(s))} r^\phi(\phi(s), a)\right] = \rho(\pi_e)$$

Consider the difference between the two variances.

$$
\begin{aligned}
D &= V^g - V^\phi \\
&= \text{Var}\left(\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)} r(s,a)\right) - \text{Var}\left(\frac{d_{\pi_e^\phi}(\phi(s))\pi_e^\phi(a|\phi(s))}{d_{\pi_\mathcal{D}^\phi}(\phi(s))\pi_\mathcal{D}^\phi(a|\phi(s))} r^\phi(\phi(s),a)\right) \\
&= \mathbb{E}_{s\sim d_{\pi_\mathcal{D}}, a\sim\pi_\mathcal{D}}\left[\left(\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)} r(s,a)\right)^2\right] - \mathbb{E}_{s\sim d_{\pi_\mathcal{D}}, a\sim\pi_\mathcal{D}}\left[\left(\frac{d_{\pi_e^\phi}(\phi(s))\pi_e^\phi(a|\phi(s))}{d_{\pi_\mathcal{D}^\phi}(\phi(s))\pi_\mathcal{D}^\phi(a|\phi(s))} r^\phi(\phi(s),a)\right)^2\right] \\
&\overset{\text{(a)}}{=} \mathbb{E}_{s\sim d_{\pi_\mathcal{D}}, a\sim\pi_\mathcal{D}}\left[\left(\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)} r(s,a)\right)^2\right] - \mathbb{E}_{s^\phi\sim d_{\pi_\mathcal{D}^\phi}, a\sim\pi_\mathcal{D}^\phi}\left[\left(\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)} r^\phi(s^\phi,a)\right)^2\right] \\
&= \sum_{s,a} d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)\left(\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)} r(s,a)\right)^2 - \sum_{s^\phi,a} d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)\left(\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)} r^\phi(s^\phi,a)\right)^2
\end{aligned}
$$

where (a) is due to Lemma 2.

We can analyze this difference by looking at one abstract state and one action and all the states that belong to it. That is, for a fixed abstract state, s^ϕ, and fixed action, a, we have:

$$
\begin{aligned}
D' &= \sum_{s\in s^\phi} d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)\left(\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)} r(s,a)\right)^2 - \left(d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)\left(\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)} r^\phi(s^\phi,a)\right)^2\right) \\
&= \left(\sum_{s\in s^\phi} \frac{(d_{\pi_e}(s)\pi_e(a|s)r(s,a))^2}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right) - \left(\frac{(d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)r^\phi(s^\phi,a))^2}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}\right) \\
&\overset{\text{(a)}}{=} r(s,a)^2\left(\left(\sum_{s\in s^\phi} \frac{(d_{\pi_e}(s)\pi_e(a|s))^2}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right) - \left(\frac{(d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)^2}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}\right)\right)
\end{aligned}
$$

where (a) is due to Definition 4 and Assumption 1.

If we can show that $D' \geq 0$ for all possible sizes of $|s^\phi|$, we will the have the original difference, D, a sum of only non-negative terms, thus proving Theorem 3. We will prove $D' \geq 0$ by proof by induction on the size of $|s^\phi|$ from 1 to some $n \leq |\mathcal{S}|$. First note that $r(s,a)^2 > 0$, so we can ignore this term.

Let our statement to prove, $P(n)$ be that $D' \geq 0$ where $n = |s^\phi|$. This is trivially true for $P(1)$ where the ground state equals the abstract state. Now consider the inductive hypothesis, $P(n)$ is true for $n \geq 1$. Now with the inductive step, we must show that $P(n+1)$ is true given $P(n)$ is true. Starting with the inductive hypothesis:

$$
D'' = \underbrace{\left(\sum_{s\in s^\phi} \frac{(d_{\pi_e}(s)\pi_e(a|s))^2}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)}\right)}_{S} - \left(\frac{\overbrace{((d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi))}^{C})^2}{\underbrace{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)}_{C'}}\right) \geq 0
$$

After making the substitutions, we have:

$$
C^2 \leq SC' \tag{6}
$$

We have the above result holding true for when the $|s^\phi| = n$. Now consider the inductive step in relation to the inductive hypothesis where a new state, s_{n+1} is added to the abstract state. We have the following difference:

$$D'' = S + \frac{\left(d_{\pi_e}(s_{n+1})\pi_e(a|s_{n+1})\right)^2}{d_{\pi_\mathcal{D}}(s_{n+1})\pi_\mathcal{D}(a|s_{n+1})} - \frac{(C + d_{\pi_e}(s_{n+1})\pi_e(a|s_{n+1}))^2}{C' + d_{\pi_\mathcal{D}}(s_{n+1})\pi_\mathcal{D}(a|s_{n+1})}$$

For ease in notation, let $x = d_{\pi_e}(s_{n+1})\pi_e(a|s_{n+1})$ and $y = d_{\pi_\mathcal{D}}(s_{n+1})\pi_\mathcal{D}(a|s_{n+1})$. The above difference is then:

$$\begin{aligned}
D'' &= S + \frac{x^2}{y} - \frac{(C+x)^2}{C'+y} \\
&= \frac{Sy + x^2}{y} - \frac{(C+x)^2}{C'+y} \\
&= \frac{1}{y(C'+y)}((Sy + x^2)(C'+y) - (C+x)^2 y) \\
&= \frac{1}{y(C'+y)}(SyC' + Sy^2 + x^2C' + x^2y - C^2y - x^2y - 2Cxy) \\
&= \frac{1}{y(C'+y)}(SyC' + Sy^2 + x^2C' - C^2y - 2Cxy)
\end{aligned}$$

The above difference, D'', is minimized most when C is as large as possible. From the inductive hypothesis, we have $C \leq \sqrt{SC'}$. The minimum difference can be written as:

$$\begin{aligned}
D'' &= \frac{1}{y(C'+y)}(Sy^2 + x^2C' - 2\sqrt{SC'}xy) \\
&= \frac{1}{y(C'+y)}(y\sqrt{S} - x\sqrt{C'})^2 \\
&\geq 0
\end{aligned}$$

So we have $D'' \geq 0$ for $|s^\phi| = n+1$, which means $D' \geq 0$. We have showed that $P(n)$ is true for all n. We now have the original difference, D, to be a sum of non-negative terms after performing this same grouping for all abstract states and actions, which results in:

$$V^\phi \leq V^g$$

$$\text{Var}\left(\frac{d_{\pi_e^\phi}(s^\phi)\pi_e^\phi(a|s^\phi)}{d_{\pi_\mathcal{D}^\phi}(s^\phi)\pi_\mathcal{D}^\phi(a|s^\phi)} r^\phi(s^\phi, a)\right) \leq \text{Var}\left(\frac{d_{\pi_e}(s)\pi_e(a|s)}{d_{\pi_\mathcal{D}}(s)\pi_\mathcal{D}(a|s)} r(s,a)\right)$$

Thus, when the covariance terms interact favorably, that is, if for any fixed $1 \leq t < k \leq T$,
$\text{Cov}\left(\frac{d_{\pi_e^\phi}(s_t^\phi,a_t)}{d_{\pi_\mathcal{D}^\phi}(s_t^\phi,a_t)} r^\phi(s_t^\phi, a_t), \frac{d_{\pi_e^\phi}(s_k^\phi,a_k)}{d_{\pi_\mathcal{D}^\phi}(s_k^\phi,a_k)} r^\phi(s_k^\phi, a_k)\right) \leq \text{Cov}\left(\frac{d_{\pi_e}(s_t,a_t)}{d_{\pi_\mathcal{D}}(s_t,a_t)} r(s_t, a_t), \frac{d_{\pi_e}(s_k,a_k)}{d_{\pi_\mathcal{D}}(s_k,a_k)} r(s_k, a_k)\right)$,
we have:

$$\text{Var}(\hat{\rho}(\pi_e^\phi)) \leq \text{Var}(\hat{\rho}(\pi_e))$$

□

A.4 Additional AbstractBestDICE Derivation Details

We proceed assuming ϕ satisfies Assumptions 1, 2, and 3 from the main text. We base the following derivation on that of DualDICE Nachum et al. (2019). The only difference between DualDICE and

BestDICE (which we use in our experiments) is that the latter enforces a positivity and unit mean constraint on the ratios.

Technical observation Consider the same technical observation made in DualDICE: the solution to the scalar convex optimization problem $\min_x J(x) \coloneqq \frac{1}{2}mx^2 - nx$, where $m \in \mathbb{R}_{>0}$ and $n \in \mathbb{R}_{\geq 0}$, is $x^* = \frac{n}{m}$. This observation can then be connected to a similar convex problem but in terms of functions where the ratio $\frac{n}{m}$ corresponds to the true abstract ratios, $d_{\pi_e^\phi}(s^\phi, a)/d_{\mathcal{D}^\phi}(s^\phi, a)$. Consider the following convex problem where $x \in \mathcal{S}^\phi \times \mathcal{A} \to [0, C] \subset \mathbb{R}$:

$$\min_{x:\mathcal{S}^\phi \times \mathcal{A} \to \mathcal{C}} J(x) := \frac{1}{2}\,\mathbb{E}_{(s^\phi,a)\sim d_{\mathcal{D}^\phi}}[x(s^\phi, a)^2] - \mathbb{E}_{(s^\phi,a)\sim d_{\pi_e^\phi}}[x(s^\phi, a)]$$

The solution to this optimization problem is the true abstract state-action ratios: $x^*(s^\phi, a) = d_{\pi_e^\phi}(s^\phi, a)/d_{\mathcal{D}^\phi}(s^\phi, a)$.

Change-of-variables As also done by Nachum et al. (2019), we next consider a change-of-variables trick to obtain an objective that can be approximated with samples from $\mathcal{D}^\phi$. However, as noted in Section 3, there are two challenges we must overcome before applying this procedure on $\mathcal{D}^\phi$. To overcome these challenges, we identified Assumptions 2 and 3 as conditions that ϕ must satisfy. With a satisfactory ϕ, we can proceed as follows. Consider an arbitrary function $\nu \in \mathcal{S}^\phi \times \mathcal{A} \to \mathbb{R}$ that satisfies $\nu(s^\phi, a) = x(s^\phi, a) + \mathcal{B}^{\pi_e^\phi}\nu(s^\phi, a)$ where $\mathcal{B}^{\pi_e^\phi}\nu(s^\phi, a) = \gamma\,\mathbb{E}_{s'^\phi \sim P^\phi(s^\phi,a), a' \sim \pi_e^\phi(s'^\phi)}[\nu(s'^\phi, a')]$. Since $x(s^\phi, a) \in [0, C]$ and $\gamma < 1$, ν is well-defined. We can the replace $x(s^\phi, a)$ with $(\nu - \mathcal{B}^{\pi_e^\phi}\nu)(s^\phi, a)$. For the second expectation, we define $\beta_t^\phi(s^\phi) := \mathbb{P}(s^\phi = s_t^\phi | s_0^\phi \sim \beta^\phi, a_k \sim \pi_e^\phi(s_k^\phi), s_{k+1}^\phi \sim P^\phi(s_k^\phi, a_k)), 0 \leq k < t$ as the abstract-state visitation probability at time-step t by the π_e^ϕ in the abstract MDP, and then have:

$$\begin{aligned}
\mathbb{E}_{(s^\phi,a)\sim d_{\pi_e^\phi}}[x(s^\phi, a)] &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\,\mathbb{E}_{s^\phi\sim\beta_t^\phi, a\sim\pi_e^\phi}[x(s^\phi, a)] \\
&= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\,\mathbb{E}_{s^\phi\sim\beta_t, a\sim\pi_e^\phi}[\nu(s^\phi, a) - \gamma\,\mathbb{E}_{s'^\phi\sim P^\phi(s^\phi,a), a'\sim\pi^\phi(s^{\phi\prime})}[\nu(s'^\phi, a')] \\
&= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\,\mathbb{E}_{s^\phi\sim\beta_t^\phi, a\sim\pi_e^\phi}[\nu(s^\phi, a)] - (1-\gamma)\sum_{t=0}^{\infty}\gamma^{t+1}\,\mathbb{E}_{s^\phi\sim\beta_{t+1}^\phi, a\sim\pi_e^\phi}[\nu(s^\phi, a)] \\
\mathbb{E}_{(s^\phi,a)\sim d_{\pi_e^\phi}}[x(s^\phi, a)] &= (1-\gamma)\,\mathbb{E}_{s^\phi\sim\beta_0^\phi, a\sim\pi_e^\phi}[\nu(s^\phi, a)]
\end{aligned}$$

where $\beta_0^\phi = d_{0^\phi}$ is the initial abstract state distribution. After the two changes, we then have:

$$\min_{\nu:\mathcal{S}^\phi\times\mathcal{A}\to\mathbb{R}} J(\nu) := \frac{1}{2}\,\mathbb{E}_{\mathcal{D}^\phi}[(\nu - \mathcal{B}^{\pi_e^\phi}\nu)(s^\phi, a)^2] - (1-\gamma)\,\mathbb{E}_{s_0^\phi\sim d_{0^\phi}, a_0\sim\pi_e^\phi}[\nu(s_0^\phi, a_0)] \tag{7}$$

where we have the optimal solution $x^*(s^\phi, a) = (\nu^* - \mathcal{B}^{\pi_e^\phi}\nu^*)(s^\phi, a) = d_{\pi_e^\phi}(s^\phi, a)/d_{\mathcal{D}^\phi}(s^\phi, a)$.

A.4.1 Optimization techniques

The optimization problem in Equation (7) presents a couple of optimization challenges. These challenges are the same as in DualDICE Nachum et al. (2019) (page 5). Namely,

- The quantity $(\nu - \nu\mathcal{B}^{\pi_e^\phi})(s^\phi, a)$ involves a conditional exepctation inside the square. When the environment dynamics are stochastic and/or the action space is large or continuous, this quantity may not be readily optimized using stochastic techniques
- Even once ν^* is determined, $(\nu^* - \nu^*\mathcal{B}^{\pi_e^\phi})(s^\phi, a)$ may not be easily computable due to the same reasons as above.

To overcome these challenges, Fenchal duality is invoked Rockafellar (1970) where for a convex function f, $f(x) = \max_\zeta x\cdot\zeta - f^*(\zeta)$, where f^* is the Fenchal conjugate of f. When $f(x) = \frac{1}{2}x^2$, $f^*(\zeta) = \frac{1}{2}\zeta^2$. Thus, the optimization given in Equation 7 becomes:

$$\min_{\nu:\mathcal{S}^\phi\times\mathcal{A}\to\mathbb{R}} J(\nu) := \frac{1}{2}\,\mathbb{E}_{\mathcal{D}^\phi}\left[\max_{\zeta}\left((\nu-\mathcal{B}^{\pi_e^\phi}\nu)(s^\phi,a)\zeta-\frac{1}{2}\zeta^2\right)\right]-(1-\gamma)\,\mathbb{E}_{s_0^\phi\sim d_{0^\phi},a_0\sim\pi_e^\phi}[\nu(s_0^\phi,a_0)] \tag{8}$$

Then by the interchangability principle Rockafellar and Wets (1998); Dai et al. (2016), we replace the inner maximization over scalars to a maximization over $\zeta : \mathcal{S}^\phi \times \mathcal{A} \to \mathbb{R}$, resulting in the optimization given by Equation (8):

$$\min_{\nu:\mathcal{S}^\phi\times\mathcal{A}\to\mathbb{R}}\max_{\zeta:\mathcal{S}^\phi\times\mathcal{A}\to\mathbb{R}} J(\nu,\zeta) := \mathbb{E}_{\mathcal{D}^\phi}[\zeta(s^\phi,a)(\nu(s^\phi,a)-\gamma\,\mathbb{E}_{a'\sim\pi_e^\phi}\nu(s'^\phi,a'))-\frac{1}{2}\zeta^2(s^\phi,a)]$$
$$-(1-\gamma)\,\mathbb{E}_{s_0^\phi\sim d_{0^\phi},a_0\sim\pi_e^\phi}[\nu(s_0^\phi,a_0)]$$

On applying the KKT conditions to the inner optimization, which is convex and quadratic, for any ν the optimal $\zeta(s^\phi,a) = (\nu-\mathcal{B}^{\pi_e^\phi}\nu)(s^\phi,a)$. Therefore, for the optimal ν^*, we have $\zeta^*(s^\phi,a) = (\nu^*-\mathcal{B}^{\pi_e^\phi}\nu^*)(s^\phi,a) = d_{\pi_e^\phi}(s^\phi,a)/d_{\mathcal{D}^\phi}(s^\phi,a)$. This derivation with DualDICE naturally extends to BestDICE since BestDICE solves the same optimization with the added constraints that $\mathbb{E}[\zeta]=1$ and $\zeta\geq 0$, which are properties we know the true ratios would satisfy.

A.5 Additional Tabular Experiments and Details

In this section, we include the remaining tabular experiments and information.

- The true value of $\rho(\pi_e)$ was determined by executing π_e for 200 episodes and averaging the results.
- In all experiments, we use $\gamma = 0.999$
- In the $\hat{d}_{\pi_e^\phi}$ vs. $d_{\pi_e^\phi}$ plots, we use batch size of 300.

In practice, Assumptions 2 and 3 may be hard to validate. To study the impact of violation of these assumptions, we consider the following variations (Figure 5) to the original TwoPath MDP presented in Figure 2. In all these MDPS, we only aggregate s_1 and s_2. The policies in the last two MDPs are the same. Figure 6 illustrates the accuracy of AbstractBestDICE's true ratio estimation. In general, we do see AbstractBestDICE fails to compute the true abstract ratios when these assumptions are violated.

A.6 Additional Function Approximation Experiments and Details

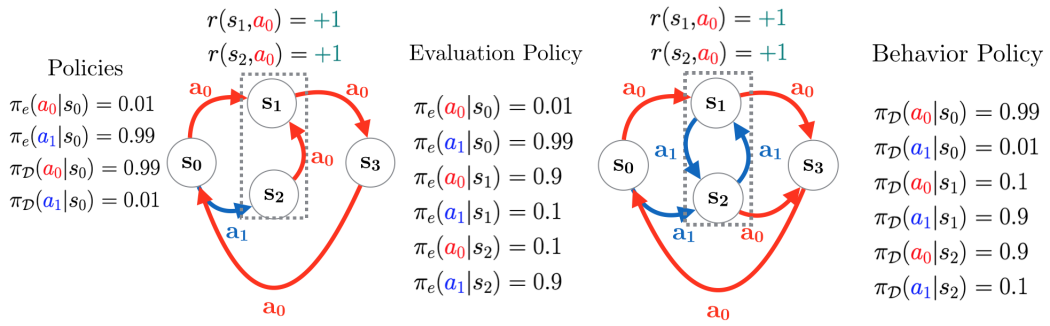
A.6.1 Oracle $\rho(\pi_e)$ Values

On each domain, we executed π_e for 200 episodes and averaged the results.

A.6.2 Policies

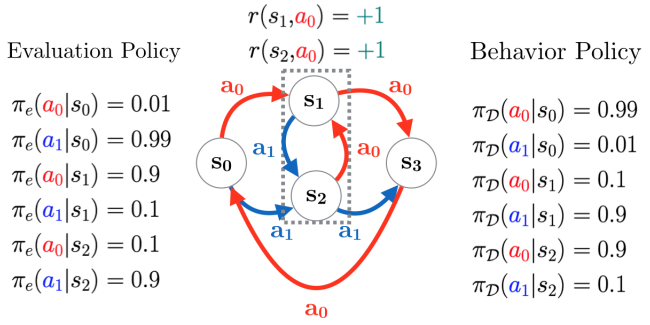
For each of the domains, we used the following policies:

- Reacher: We trained a policy using PPO Schulman et al. (2017). π_e was the trained policy after 100k time-steps with a standard deviation of 0.1 on the action dimensions while π_b used 0.5 as the standard deviation.
- Walker2D: We trained a policy using PPO Schulman et al. (2017). π_e was the trained policy after 100k time-steps with a standard deviation of 0.1 on the action dimensions while π_b used 0.5 as the standard deviation.
- Pusher: We trained a policy using PPO Schulman et al. (2017). π_e was the trained policy after 100k time-steps with a standard deviation of 0.1 on the action dimensions while π_b used 0.5 as the standard deviation.
- AntUMaze: We used the policies made available Fu et al. (2021). π_e was the final 10th snapshot saved and π_b was the 5th snapshot. Each also had 0.1 standard deviation on the action dimensions.



(a) Assumption 2 (transition dynamics) violated

(b) Assumption 3 (π_e action equality) violated



(c) Assumptions 2 and 3 violated. (a_0 and a_1 in s_2 switched)

Figure 5: Variations to the TwoPath MDP from Figure 2 to violate assumptions.

A.6.3 Trajectory Length

For each of the domains, the trajectory length is: 200 for Reacher, 500 for Walker2D, 300 for Pusher, and 500 for AntUMaze.

A.6.4 Hyperparameters

For BestDICE and AbstractBestDICE, we fixed the following hyperparameters:

- $\gamma = 0.995$ in all experiments.
- Neural net architecture: All neural networks are 2 layers with 64 hidden units using tanh activation.
- Unit mean constraint learning rate Zhang et al. (2020b); Yang et al. (2020): $\lambda = 1e^{-3}$.
- Optimizer: Adam optimizer with default parameters in Pytorch.
- Positivity constraint: squaring function on the last layer of the neural network.

We conducted a search for the learning rate of ν (α_ν) and learning rate of ζ (α_ζ), . The learning rate search for $(\alpha_\nu, \alpha_\zeta)$ was over $\{(5e^{-5}, 5e^{-5}), (1e^{-4}, 1e^{-4}), (3e^{-4}, 3e^{-4}), (7e^{-4}, 7e^{-4}), (1e^{-3}, 1e^{-3})\}$. The optimal hyperparameters ($\alpha_\nu = \alpha_\zeta$) for each environment and batch size were:

	5	10	50	75	100	300	500	1000
Reacher	$5e^{-5}$	$1e^{-4}$	$1e^{-3}$	$1e^{-4}$	$1e^{-4}$	$7e^{-4}$	$1e^{-3}$	$7e^{-4}$
Walker2D	$1e^{-3}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$
Pusher	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$	$1e^{-4}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$
AntUMaze	$3e^{-4}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$

Table 1: Optimal hyparameters for AbstractBestDICE on each batch size and environment.

	5	10	50	75	100	300	500	1000
Reacher	$5e^{-5}$	$1e^{-4}$	$5e^{-5}$	$1e^{-4}$	$5e^{-5}$	$1e^{-4}$	$3e^{-4}$	$1e^{-3}$
Walker2D	$5e^{-5}$	$3e^{-4}$	$7e^{-4}$	$7e^{-4}$	$7e^{-4}$	$1e^{-4}$	$3e^{-4}$	$1e^{-4}$
Pusher	$5e^{-5}$	$1e^{-4}$	$1e^{-4}$	$5e^{-5}$	$1e^{-4}$	$5e^{-5}$	$1e^{-4}$	$5e^{-5}$
AntUMaze	$1e^{-4}$	$3e^{-4}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$	$5e^{-5}$

Table 2: Optimal hyperparameters for BestDICE on each batch size and environment.

A.6.5 Empirical Estimator

In practice we use a weighted importance sampling Yang et al. (2020) approach for the function approximation cases to estimate $\rho(\pi_e)$ (same for BestDICE):

$$\hat{\rho}(\pi_e^\phi) = \frac{\sum_{i=1}^{N} \frac{d_{\pi_e^\phi}(s_i^\phi, a_i)}{d_{\pi_\mathcal{D}^\phi}(s_i^\phi, a_i)} r^\phi(s_i^\phi, a_i)}{\sum_{i=1}^{N} \frac{d_{\pi_e^\phi}(s_i^\phi, a_i)}{d_{\pi_\mathcal{D}^\phi}(s_i^\phi, a_i)}}$$

A.6.6 Misc Abstraction Details

- For Walker2D, we modified the default reward function from incremental distance covered at each time-step to distance from start location at each time-step to ensure Assumption 1 is satisfied.
- For AntUMaze, the reward function is originally $r(s')$ i.e. it is based on the *next* state that the ant moves to. To ensure Assumption 1 is satisfied, we changed this reward function to be of the *current* state, $r(s)$.

A.6.7 Additional Results

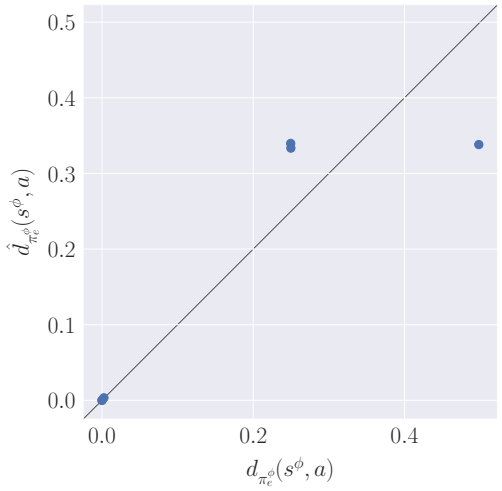
Additional Hyperparameter Robustness Results In general, we can see AbstractBestDICE can be much more robust than BestDICE to hyperparameter tuning.

Training Stability In Figure 8 we show that ϕ can improve training stability.

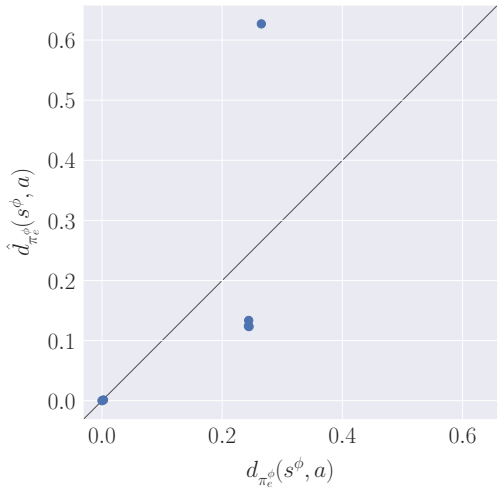
Abstract Quality and Data-Efficiency. We find that not all abstractions that satisfy Assumption 1 lead to better performance. For example, the following are valid abstractions on the Reacher task: 1) the Euclidean distance between the arm and goal, $s^\phi \in \mathbb{R}$ and the 3D vector between the arm and goal, $s^\phi \in \mathbb{R}^3$ (Figure 9). However, in practice we found that these were unreliable. One possible reason for this unreliability is that these abstractions are incredibly extreme and the algorithm may be unable to differentiate between abstract state, resulting in outputting similar $\zeta^\phi(s^\phi, a) \forall s^\phi$.

A.7 Hardware For Experiments

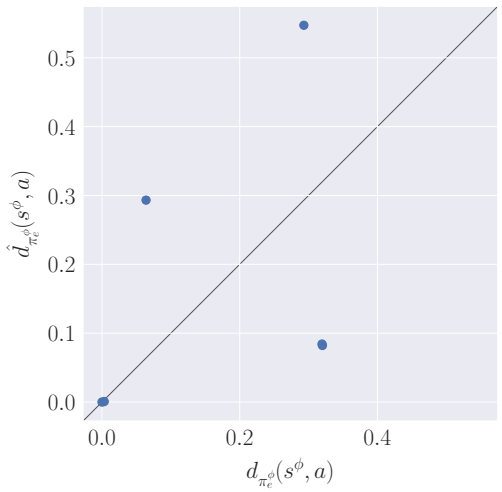
- Distributed cluster on HTCondor framework
- Intel(R) Xeon(R) CPU E5-2470 0 @ 2.30GHz
- RAM: 5GB
- Disk space: 4GB



(a) Assumption 2 (transition dynamics) violated

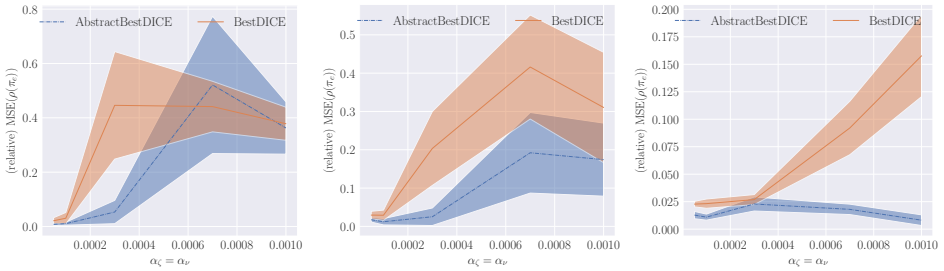


(b) Assumption 3 (π_e action equality) violated

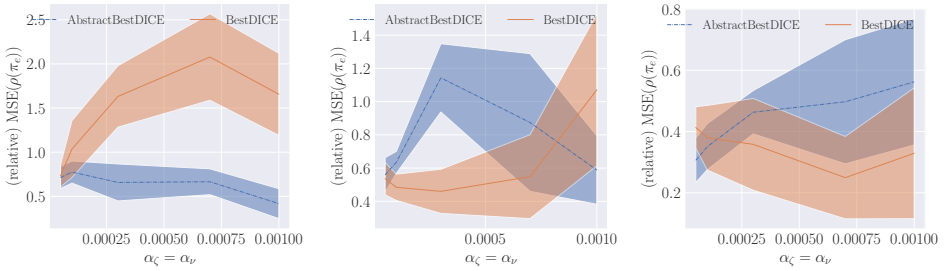


(c) Assumptions 2 and 3 violated

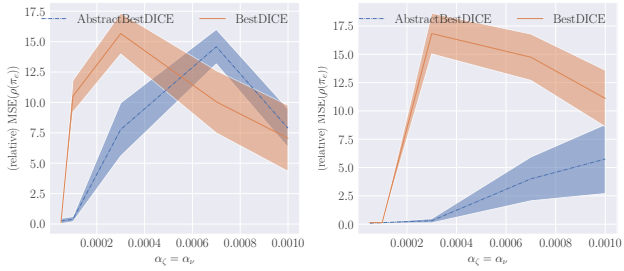
Figure 6: True ratio estimations with assumption violations. Closer to the linear line is better.



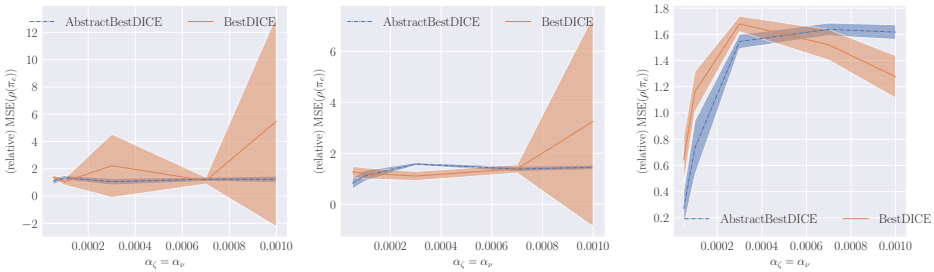
(a) Reacher (Batch size: 5) (b) Reacher (Batch size: 10) (c) Reacher (Batch size: 50)



(d) Walker2D (Batch size: 5) (e) Walker2D (Batch size: 10) (f) Walker2D (Batch size: 50)

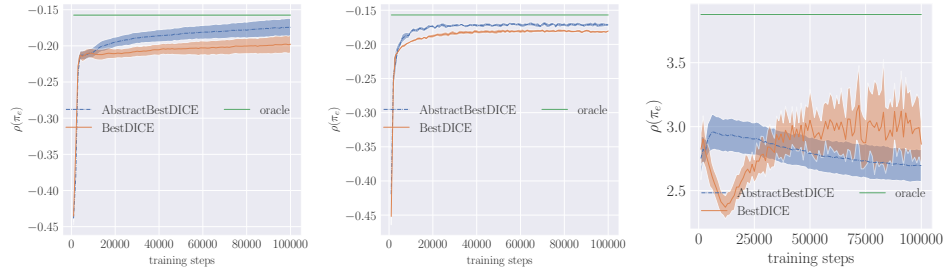


(g) Pusher (Batch size: 5) (h) Pusher (Batch size: 10)

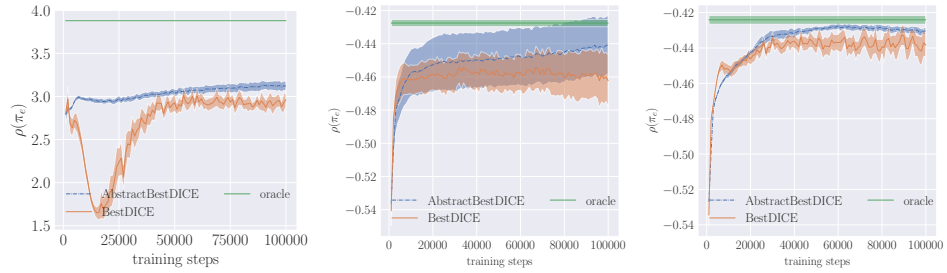


(i) AntUMaze (Batch size: 5) (j) AntUMaze (Batch size: 10) (k) AntUMaze (Batch size: 50)

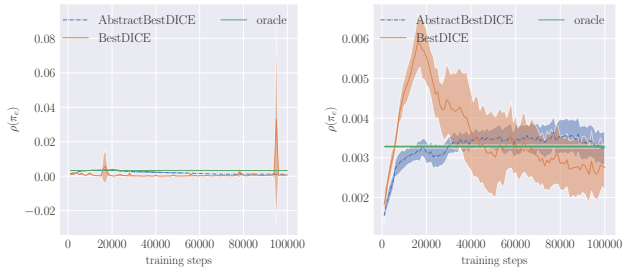
Figure 7: Hyperparameter sensitivity graph for BestDICE and AbstractBestDICE. $\alpha_\zeta = \alpha_\nu$ Errors are computed over 15 trials with 95% confidence intervals. Lower is better. Pusher for batch size of 50 is shown in the main paper.



(a) Reacher (Batch size: 10) (b) Reacher (Batch size: 500) (c) Walker2D (Batch size: 10)

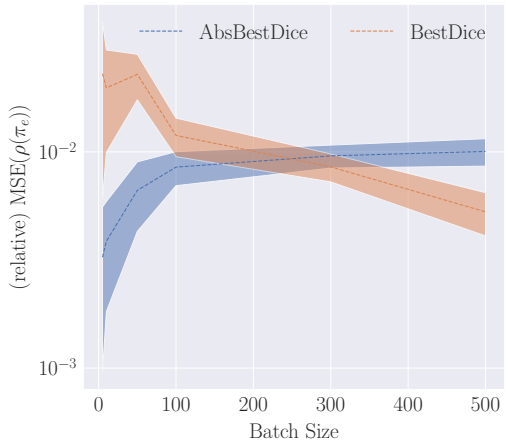


(d) Walker2D (Batch size: 500) (e) Pusher (Batch size: 10) (f) Pusher (Batch size: 500)



(g) AntUMaze (Batch size: 10) (h) AntUMaze (Batch size: 500)

Figure 8: Reward vs. Training Steps. Errors are computed over 15 trials with 95% confidence intervals. These figures illustrate the training stability of AbstractBestDICE over BestDICE. Lower is better.



(a) Reacher

Figure 9: Relative MSE vs. Batch Size. y axis is log-scaled. Errors are computed over 15 trials with 95% confidence intervals. This figures illustrate valid abstractions can be more data-inefficient than the ground equivalents. Lower is better.