AVG-DICE: Stationary Distribution Correction by Regression

Anonymous authors
Paper under double-blind review

Keywords: Off-Policy Evaluation, State Distribution Correction, Importance Sampling, Distribution Shift.

Summary

Off-policy policy evaluation (OPE), an essential component of reinforcement learning, has long suffered from stationary state distribution mismatch, undermining both stability and accuracy of OPE estimates. While existing methods correct distribution shifts by estimating density ratios, they often rely on expensive optimization or backward Bellman-based updates and struggle to outperform simpler baselines. We introduce AVG-DICE, a computationally simple Monte Carlo estimator for the density ratio that averages discounted importance sampling ratios, providing an unbiased and consistent correction. AVG-DICE extends naturally to non-linear function approximation using regression, which we roughly tune and test on OPE tasks based on Mujoco Gym environments and compare with state-of-the-art density-ratio estimators using their reported hyperparameters. In our experiments, AVG-DICE is at least as accurate as state-of-the-art estimators and sometimes offers orders-of-magnitude improvements. However, a sensitivity analysis shows that best-performing hyperparameters may vary substantially across different discount factors, so a re-tuning is suggested.

Contribution(s)

1. We reformulate the state distribution ratio between the discounted stationary distribution of the target policy and the undiscounted stationary distribution of the behaviour policy as a new consistent estimator, leveraging a dataset collected under the behaviour policy.
2. We show that this consistent estimator corrects state distribution shifts in off-policy data, and reweighting each data point with our estimator provides an unbiased estimate for any function.
3. We introduce AVG-DICE, an algorithm that estimates density ratios via regression.
4. We prove the convergence of our update rules under linear function approximation.
5. We evaluate AVG-DICE against prior algorithms and demonstrate that it achieves either dominant or comparable performance across all baselines.
   Context: Our algorithm is sensitive to changes in the discount factor, and we recommend re-tuning it for each discount setting to ensure optimal performance.

AVG-DICE: Stationary Distribution Correction by Regression

Anonymous authors
Paper under double-blind review

Abstract

Off-policy policy evaluation (OPE), an essential component of reinforcement learning, has long suffered from stationary state distribution mismatch, undermining both stability and accuracy of OPE estimates. While existing methods correct distribution shifts by estimating density ratios, they often rely on expensive optimization or backward Bellman-based updates and struggle to outperform simpler baselines. We introduce AVG-DICE, a computationally simple Monte Carlo estimator for the density ratio that averages discounted importance sampling ratios, providing an unbiased and consistent correction. AVG-DICE extends naturally to nonlinear function approximation using regression, which we roughly tune and test on OPE tasks based on Mujoco Gym environments and compare with state-of-the-art density-ratio estimators using their reported hyperparameters. In our experiments, AVG-DICE is at least as accurate as state-of-the-art estimators and sometimes offers orders-of-magnitude improvements. However, a sensitivity analysis shows that best-performing hyperparameters may vary substantially across different discount factors, so a re-tuning is suggested.

1 Introduction

Off-policy evaluation (OPE) aims to estimate the expected cumulative return of a target policy using data collected from a different behaviour policy. Assessing a policy with pre-collected data before deployment is crucial, as executing an unqualified policy can lead to undesirable consequences (Levine et al. 2020), including life-threatening risks in applications such as surgical robotics and self-driving vehicles. A straightforward approach is to directly average the observed rewards. However, distribution shift introduces bias into value estimation, and even temporal difference methods cannot provide an unbiased evaluation of the target policy under such shifts (Sutton et al. 2016).

A common approach to addressing distribution shift is importance sampling (IS) (Precup et al. 2001), which reweights samples based on the ratio between two distributions to provide an unbiased estimation. However, when correcting cumulative returns along a trajectory, IS requires multiplying these ratios over multiple steps, leading to high variance — a problem known as the curse of the horizon. To mitigate this, researchers have explored marginalized IS ratios for stationary state distributions (Liu et al. 2018). Current estimators leverage the recursive property of stationary distributions to formulate optimization tasks. This recursion results in a backward Bellman-based update, where the value at the next step depends on the current step's value (Hallak & Mannor 2017). However, the expectation in the backward Bellman recursion cannot be unbiasedly evaluated without double sampling. Moreover, off-policy Bellman updates with function approximation, known as the deadly triad (Baird 1995), are prone to instability and typically lack convergence guarantees.

Several studies have proposed novel optimization frameworks, such as primal-dual optimization (Nachum et al. 2019) or multistage optimization (Uehara et al. 2020), to avoid directly minimizing the backward Bellman error. The optimal solutions corresponding to their novel losses equal to the desired state density ratio, proven in the tabular case (Liu et al. 2019). However, these methods introduce additional complexity to the estimation process. This raises an important question: Can we revisit Monte Carlo methods to develop a new estimator that is both theoretically sound and computationally simpler? Previously, a state distribution corrector based on the Monte Carlo expansion of the stationary distribution was developed to account for the missing discount factor in stationary state distributions for policy gradient algorithms (Che et al. 2023). However, the idea was not explored in the off-policy setting.

In this paper, we propose a novel estimator for the stationary state distribution ratio, called the average state distribution correction estimation (AVG-DICE). It leverages the Monte Carlo expansion rather than the recursive property used in prior approaches. Our approach computes the average of all discounted importance sampling ratio products corresponding to a given state in the dataset. We prove that this method provides a consistent estimation of the density ratio between the discounted target and the undiscounted behaviour stationary state distributions. Also, it gives an unbiased estimation of any function.

Furthermore, our estimator can be learned via a least squares regression task, offering a simple and effective approach to approximating the state distribution ratio. In the case of linear function approximation, we establish its asymptotic convergence to the same fixed point as minimizing the mean squared error with the exact density ratio, under standard assumptions used in temporal difference (TD) convergence analysis (Yu 2015).

To evaluate our estimator, we conduct experiments on several discrete classic control tasks and continuous MuJoCo tasks (Todorov et al. 2012), using a pre-collected fixed off-policy dataset with batch updates. Our estimator achieves dominant performance on most tasks when appropriately tuned for the required discount factor and remains competitive on others. Additionally, it demonstrates the fastest convergence to a stable value, making it practical for integration into other algorithms. However, our algorithm is sensitive to changes in the discount factor, and we recommend re-tuning it for each discount setting to ensure optimal performance.

2 Background

Notation We let $\Delta(\mathcal{X})$ denote the set of probability distributions over a finite set $\mathcal{X}$. Let $\mathbb{R}$ denote the set of real numbers, $\mathbb{N}$ be the set of non-negative integers, $\mathbb{N}^+$ be the set of positive integers, and $\mathbb{1}$ be the indicator function.

Markov Decision Process We consider finite Markov decision process (MDP) (Sutton & Barto 2018) defined by a tuple $M = \langle \mathcal{S} \cup \{\varsigma\}, \mathcal{A}, r, P, \nu, \gamma \rangle$, where $\mathcal{S}$ is a finite state space, ς is a termination state, $\mathcal{A}$ is the action space, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition matrix, $\nu \in \Delta(\mathcal{S})$ is the distribution of the initial state, and $\gamma \in (0, 1)$ is the discount factor. At each step j, the agent applies an action A_j sampling from a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ at state S_j. Then, the agent receives a reward R_j and transits to the next state S'_j. Our paper focuses on episodic tasks, where a trajectory, denoted by $\tau = \{S_j, A_j, R_j, S'_j\}_{j=0}^{T-1}$, ends at the termination state ς at step T. We define the random variable $T \in \mathbb{N}^+$ as *the length of the trajectory*.

The agent's goal is to evaluate a policy π by estimating the expected discounted cumulative returns. The expected discounted cumulative return, denoted by $J(\pi)$, is defined as

$$J(\pi) = (1 - \gamma)\mathbb{E}_\pi \left[\sum_{j=0}^{\infty} \gamma^j r(S_j, A_j) \right],$$

where we use $\mathbb{E}_\pi$ to denote the expectation under the distribution induced by π and the environment. Notice that after reaching the termination state, the probability of any state showing up is zero, that is, $\mathbb{P}_\pi(S_j = s) = 0\,, \forall s \in \mathcal{S}$, if $T \leq j$ and rewards also equal zero.

The *Q-value* represents the expected cumulative rewards starting from a state-action pair (s, a) following a policy π, defined as

$$q_\pi(s,a) = \mathbb{E}_\pi \left[\sum_{j=0}^{\infty} \gamma^j r(S_j, A_j) \middle| S_0 = s, A_0 = a \right] . \tag{1}$$

Off-Policy Evaluation We consider evaluating a target policy π using a dataset $\mathcal{D}$ that consists of K trajectories as $\{\tau_i\}_{i=1}^K = \left\{ \left\{ S_j^i, A_j^i, R_j^i, (S_j')^i) \right\}_{j=0}^{T^i-1} \right\}_{i=1}^K$. When trajectories are collected under a *behaviour policy*, denoted by μ, differing from the target policy, we call the learning off-policy.

To correct this distribution shift, the *importance sampling (IS) ratio* $\rho(a|s) = \frac{\pi(a|s)}{\mu(a|s)}$ is often used to obtain unbiased estimators. The *IS-ratio product*, denoted by $\rho_{0:j-1} = \prod_{k=0}^{j-1} \rho(A_k|S_k)$, adjusts the distribution of an entire trajectory, $\{S_0, A_0, \cdots, S_j\}$, from the behaviour policy to the target one. Notice the product is initialized at one, denoted as $\rho_{0:-1} = 1$, with some abuse of notations.

This dataset can also be expressed in terms of individual transitions as $\mathcal{D} = \{(S_t, A_t, R_t, S_t', \text{time}_t, \rho_{\text{prod},t})\}_{t=0}^{n-1}$ where n is the dataset size, t is the index for each transition, time_t represents the step of s_t in its trajectory and $\rho_{\text{prod},t}$ for the corresponding IS products $\rho_{0:\text{time}_t-1}$ until s_t. To further simplify notation, let I_s indicate the set of step t such that $S_t = s$.

Off-policy TD estimates Q-values by $\hat{q}_\theta$ and takes a semi-gradient of the empirical temporal difference errors, which equals

$$\min \mathcal{L}(\theta; \mathcal{D}) = \frac{1}{n} \sum_{t=0}^{n-1} \left(R_t + \gamma \hat{q}_\theta \left(S_t', A_t' \right) - \hat{q}_\theta(S_t, A_t) \right)^2 , \tag{2}$$

where the next action used for the bootstrapping target is sampled from the target policy, that is, $A_t' \sim \pi(\cdot|S_t')$. TD evaluates the target policy by expected value estimation of initial state-action pairs, that is, $(1-\gamma)\mathbb{E}_{S_0 \sim \nu, A_0 \sim \pi(\cdot|S_0)} [\hat{q}_\theta(S_0, A_0)]$. However, the dataset's state distribution shift is not corrected from the behaviour policy to the target policy, leading to bias in the estimation.

Irreducible Markov Chain The Markov decision process under a policy π forms a Markov chain, denoted by $\langle \mathcal{S} \cup \{\varsigma\}, P_\pi \rangle$, where $P_\pi(s'|s) = \sum_{a \in \mathcal{A}} \pi(a|s) P(s'|s, a)$ for any $s, s' \in \mathcal{S}$. A Markov chain is said to be *irreducible*, if for any two states, s and s', the probability of transiting between these two states is positive at some time step, that is, $P_\pi(S_j = s' \text{ for some } j > 0 | S_0 = s) > 0$ and $P_\pi(S_j = s \text{ for some } j > 0 | S_0 = s') > 0$.

The *recurrence time* of a state s, denoted by $\tau_s^+(s)$, is defined as the time elapsed to revisit a state s, that is, $\tau_s^+(s) = \min\{j > 0 : S_j = s, S_0 = s\}$. A *positive recurrent* state has a finite expected recurrence time, that is, $\mathbb{E}_\pi[\tau_s^+(s)] < \infty$. Note that no assumptions are made in the background section; the terms irreducibility and positive recurrence are presented solely for later use.

Stationary State Distribution The discounted stationary state distribution, denoted by $d_{\pi,\gamma}$, is defined as the distribution satisfying the following equation for all states $s' \in \mathcal{S}$:

$$\sum_{s \in \mathcal{S}} d_{\pi,\gamma}(s)[\gamma P_\pi(s'|s) + (1-\gamma)\nu(s')] = d_{\pi,\gamma}(s'). \tag{3}$$

A common analytical form of the discounted stationary distribution can be written as

$$d_{\pi,\gamma}(s) = (1-\gamma) \sum_{j=0}^{\infty} \gamma^j \mathbb{P}_\pi(S_j = s). \tag{4}$$

Given the form of the discounted stationary distribution, the expected discounted cumulative return, $J(\pi)$, can also be written as

$$J(\pi) = \sum_{s \in \mathcal{S}} d_{\pi,\gamma}(s) r_\pi(s), \tag{5}$$

where $r_\pi(s) = \mathbb{E}_{A \sim \pi(\cdot|s)}[r(s, A)]$.

The undiscounted stationary distribution, denoted by d_π is defined as the distribution satisfying $\sum_{s\in\mathcal{S}} d_\pi(s)P_\pi(s'|s) = d_\pi(s')$. This distribution is also regarded as the limiting distribution of state-action visitation at each step. However, in episodic tasks, the step count does not approach infinity. To define a limit distribution in this setting, the trajectory is considered to restart from the initial state distribution upon termination. Note that this restart does not occur in practice and is introduced purely for definitional purposes.

Repeating the transition n steps can give K terminated trajectory and one incomplete trajectory, denoted by $\left\{\left\{(S_j^i, A_j^i, R_j^i, (S_j')^i)\right\}_{j=0}^{T^i-1}\right\}_{i=1}^{K} \cup \left\{(S_j^{K+1}, A_j^{K+1}, R_j^{K+1}, (S_j')^{K+1})\right\}_{j=0}^{n-\sum_{k=1}^{K} T^k}$. We relabel the transition by t as $\{(S_t, A_t, R_t, S_t')\}_{t=0}^{n-1}$. The undiscounted stationary distribution has multiple analytical forms stated in Sutton and Barto (2018) and Grimmett and Stirzaker (2020, Theorem 6.4.3), summarized in a lemma from (Che et al. 2023).

Lemma 2.1 (Forms of Undiscounted Stationary Distribution)**.** *Under the irreducibility of the Markov chain and positive recurrences of all states under all policies π, we have the following:*

$$d_\pi(s) = \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathbb{P}_\pi(S_t = s) = \frac{1}{\mathbb{E}_\pi[\tau_s^+(s)]}. \tag{6}$$

3 Related Works

TD with linear function approximation converges when data is sampled as trajectories under the target policy (Tsitsiklis & Van Roy 1996), but linear TD with off-policy state distribution is not guaranteed to converge (Che et al. 2024). This issue is called the deadly triad. Meanwhile, the policy estimation is biased under the state distribution shift.

The data distribution can be corrected by importance sampling (Precup 2000, Precup et al. 2001). However, these approaches suffer from high variance when correcting the distributions of trajectories with products of IS ratios. Later papers work on estimating state distribution ratios to avoid the ratio product (Hallak & Mannor 2017, Yang et al. 2020, Fujimoto et al. 2021).

The state distribution ratio can be estimated based on the backward recursion for the stationary distribution shown in Equation 3. A backward Bellman recursion for the density ratio $w(s)$ can then be built for all state s', and the temporal difference error for the density ratio estimator, denoted by $TD(s')$, is defined as

$$TD(s') := \mathbb{E}_{(S,A,S')\sim d_\mu}\left[-w(S') + \gamma w(S)\frac{\pi(A|S)}{\mu(A|S)}|S' = s'\right] + (1-\gamma)\rho(s').$$

This temporal error equals zero, if $w(s) = \frac{d_{\pi,\gamma}(s)}{d_\mu(s)}$ provided that non-zero target policy $\pi(a|s) > 0$ implying the behaviour policy being non-zero $\mu(a|s) > 0$ for all state-action pairs (Nachum et al. 2019). COP-TD (Hallak & Mannor 2017, Gelada & Bellemare 2019) minimizes the above temporal difference (TD) error. However, a backward TD estimate cannot be unbiasedly computed from a dataset without double sampling unless the behaviour policy is concentrated on a single state. Meanwhile, the algorithm lacks a convergence guarantee.

Several other works (Liu et al. 2018- 2019, Uehara et al. 2020) design novel loss functions based on the recursive properties of the state distribution instead of directly minimizing the TD error. These losses reach zero if and only if the solution is the density ratio, providing new multi-stage optimization objectives for ratio approximation. On the other hand, DualDice (Nachum et al. 2019) introduces a primal-dual optimization framework by reformulating the problem with the Fenchel conjugate. GenDice (Zhang et al. 2020a) estimates the density ratio $w(s)$ and minimizes the f-divergence between the estimated and true stationary distributions $d_\mu(s)w(s)$ and $d_{\pi,\gamma}(s)$, showing greater stability than DualDice for high discount factors but lacking convex-concavity. However, these multi-stage or primal-dual optimization techniques lack the convergence guarantee, and the training is less stable with multiple variables.

GradientDice (Zhang et al. 2020b) replaces f-divergence in GenDice with a weighted L2-norm, ensuring convex-concave and convergence properties under linear function approximation. BestDice (Yang et al. 2020) unifies these multi-stage and primal-dual methods into a general objective, identifying optimal regularization choices in their BestDice algorithm. However, the learning stability still needs improvement.

Successor Representation Distribution Correction Estimation (SR-DICE) (Fujimoto et al. 2021) builds on successor features and derives a loss equivalent to minimizing the mean squared error to the density ratio under linear function approximation. It achieves lower policy evaluation errors than other density estimators but still underperforms deep off-policy TD. Meanwhile, state-action representation features and successor features require pre-training, introducing additional approximation errors and increasing computation.

4 Distribution Corrector

We derive a novel expression for the state density ratio, leading to a consistent estimator. This estimator computes the average of discounted IS-ratio products for each state using an off-policy dataset. Our algorithm, AVG-DICE, is named for its averaging approach in approximating this estimator. As the dataset size approaches infinity, our estimator converges to the true density ratio. Meanwhile, it corrects the distribution shift from the dataset's sampling distribution to the target policy's discounted stationary distribution, consequently providing an unbiased estimate for any function by reweighting each state by our estimator.

Recall that a dataset consists of K trajectories and is presented as $\mathcal{D} = \{(S_t, A_t, R_t, S'_t, \text{time}_t, \rho_{\text{prod},t})\}_{t=0}^{n-1}$, where time_t represents the step of S_t in its trajectory and $\rho_{\text{prod},t}$ for the corresponding IS products until S_t. I_s indicates the set of label t such that $S_t = s$.

We first assume the following necessary condition for applying marginalized importance sampling.

Assumption 4.1. If $d_{\pi,\gamma} > 0$, then $d_\mu > 0$.

This assumption is made for all distribution correction estimators (Yu 2015, Zhang et al. 2020b), requiring the off-policy distribution to cover the target distribution.

Also, for episodic tasks, it is normal to consider the trajectory length to have a finite expectation.

Assumption 4.2.

$$\mathbb{E}_\mu[T] < \infty.$$

Now we are ready to present the novel formulation of the density ratio.

Proposition 4.3. *[Consistency] Given*

- *a finite Markov decision process,*
- *a dataset $\mathcal{D}$ collected under a behaviour policy μ, and*
- *a target policy π*

such that Assumption 4.1 and 4.2 are satisfied, then for state s with

$$d_{\pi,\gamma}(s) > 0,$$

we have the density ratio equal

$$\frac{d_{\pi,\gamma}(s)}{d_\mu(s)} = \lim_{n\to\infty} \frac{n}{K}(1-\gamma)\mathbb{E}_{t\sim I_s}[\gamma^{time_t}\rho_{prod,t}], \tag{7}$$

where n is the number of transitions, and K denotes the number of trajectories.

We derive a consistent estimator based on the above proposition, defined as

$$c_{\mathcal{D}}(s) = \frac{n}{K}(1-\gamma)\mathbb{E}_{t\sim I_s}[\gamma^{\text{time}_t}\rho_{\text{prod},t}]. \tag{8}$$

The expectation is taken over step where state s appears and can be expressed as

$$\mathbb{E}_{t\sim I_s}[\gamma^{\text{time}_t}\rho_{\text{prod},t}] = \frac{\sum_{t=0}^{n-1}\gamma^{\text{time}_t}\rho_{\text{prod},t}\mathbb{1}[S_t = s]}{\sum_{t=0}^{n-1}\mathbb{1}[S_t = s]}, \tag{9}$$

where $\mathbb{1}[S_t = s]$ equals to one if s appears at step t. The denominator counts the number of times state s occurs in the dataset, while the numerator sums the corresponding discounted IS-ratio products. Thus, the expectation in our estimator effectively averages all discounted IS products associated with state s.

Our main theorem shows that reweighting each data by our estimator gives an unbiased estimator for any function.

Theorem 4.4 (Unbiasedness)**.** *Given*

- *a finite Markov decision process,*
- *a dataset $\mathcal{D}$ collected under a behaviour policy μ, and*
- *a target policy π*

such that Assumption 4.1 and 4.2 are satisfied, reweighting data by our average correction gives unbiased estimation for any function $f : \mathcal{S} \to \mathbb{R}$, that is,

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{S\sim\mathcal{D}}\left[c_{\mathcal{D}}(S)f(S)\right]\right] = \mathbb{E}_{S\sim d_{\pi,\gamma}}[f(S)], \tag{10}$$

where $\mathbb{E}_{\mathcal{D}}$ means expectation over trajectories sampled under the behaviour policy, and $\mathbb{E}_{S\sim\mathcal{D}}$ representing sampling states uniformly from the dataset.

This theorem holds because our estimator equals the ratio of an unbiased and consistent estimation of the discounted target distribution, denoted as $\hat{d}_{\pi,\gamma}(s) = \frac{1}{K}\sum_{i=1}^{K}\sum_{j\geq 0}\gamma^j\rho^i_{0:j-1}\mathbb{1}[S^i_j = s]$ to the sampling distribution from the dataset, denoted as $\hat{d}(s) = \dfrac{\sum_{t=0}^{n-1}\mathbb{1}[S_t = s]}{n}$. Therefore, as long as we can calculate our estimator, the distribution shift can be solved.

5 AVG-DICE Algorithm

In this section, we work on how to evaluate our derived estimator $c_{\mathcal{D}}(s)$. Our estimator averages the corresponding discounted IS-ratio products for a state s. However, in high-dimensional state spaces, direct averaging by state counting is infeasible. Thus, we propose to approximate the expectation of discounted IS-ratio products via regression.

We first introduce our regression losses and propose the AVG-DICE algorithm. Then, we show that with linear function approximation, incrementally updating our loss results in a convergent algorithm. The fixed point of this update corresponds to minimizing the mean squared error (MSE) to the true density ratio with regularization.

5.1 Loss

We learn our estimator as a ratio model by minimizing the least squares error. Solving least squares regression with Markovian data is well studied but generally requires more samples compared to i.i.d. learning tasks (Nagaraj et al. 2020). In our setup, the ratio model takes states as inputs and is trained by minimizing the mean squared error between its output $f_\theta(s_t)$ and its corresponding regression target $\gamma^{\text{time}_t}\rho_{\text{prod},t}$. In this case, the ratio is estimated by $\frac{n}{K}(1-\gamma)f_\theta(s)$. The expected discounted cumulative return can be estimated by

$$\hat{J}(\pi) = \frac{1}{n}\sum_{t=0}^{n-1}\frac{n}{K}(1-\gamma)f_\theta(S_t)R_t. \tag{11}$$

In the regression with a fixed dataset, a parameter regularization is usually added to avoid overfitting. Our algorithm also uses $\frac{\lambda_1\|\theta\|_2^2}{2}$ to regularize, where λ_1 is the regularization parameter.

Meanwhile, same as GradientDice, the learnt ratio should ensure that $\sum_s d_\mu(s)\frac{n}{K}(1-\gamma)f_\theta(s) \approx \sum_s d_\mu(s)\frac{d_{\pi,\gamma(s)}}{d_\mu(s)} = 1$. So our algorithm further regularizes by the loss $\frac{\lambda_2}{2}(\sum_s d_\mu(s)\frac{n}{K}(1-\gamma)f_\theta(s)-1)^2$, called the *distribution regularization*. An expectation in a square loss cannot be estimated unbiasedly using samples. Thus, this regularization term is re-written by the Fenchel conjugate as

$$\lambda_2(\max_{\eta\in\mathbb{R}} \mathbb{E}_{s\sim d_\mu}[\eta\frac{n}{K}(1-\gamma)f_\theta(s)-\eta] - \frac{\eta^2}{2}). \tag{12}$$

The loss given a dataset $\mathcal{D}$ is written as

$$\begin{aligned}\min_\theta \mathcal{L}(\theta;\mathcal{D}) := \mathbb{E}_{S_t\sim\mathcal{D}}\left[\frac{1}{2}(f_\theta(S_t)-\gamma_t^{\text{time}}\rho_{\text{prod},t})^2\right] + \frac{\lambda_1\|\theta\|_2^2}{2} \\ + \lambda_2\left(\max_{\eta\in\mathbb{R}} \mathbb{E}_{S_t\sim\mathcal{D}}[\eta\frac{n}{K}(1-\gamma)f_\theta(S_t)-\eta] - \frac{\eta^2}{2}\right).\end{aligned} \tag{13}$$

5.2 Convergence Analysis

This section focuses on the linear function approximation with $f_\theta(s) = \phi(s)^\top\theta$, where $\phi(s)\in\mathbb{R}^d$ is a given state feature and $\theta\in\mathbb{R}^d$ is the parameter. We denote $\Phi\in\mathbb{R}^{|\mathcal{S}|\times d}$ as the feature matrix, where each row corresponds to the feature vector of a particular state s.

At each step t, the agent takes an action according to the behaviour policy at state s_t. If the trajectory terminates, the agent restarts according to the initial distribution. The algorithm updates the parameters θ and η in our distribution regularization at each step with the newly collected transition following our loss shown in Equation 13. The regression target, denoted by $y_t = \gamma_t^{\text{time}}\rho_{\text{prod},t}$, equals to the discounted IS-ratio products computing using the state's current trajectory, where time_t represents the step of the state in its current trajectory and $\rho_{\text{prod},t} = \rho_{0:\text{time}_t-1}$.

Instead of using a running scalar of $\frac{t}{K}$ in the loss, we evaluate an average trajectory length H at the beginning and keep it fixed. This fixed multiplier simplifies the proof. We hypothesize that using the original one $\frac{t}{K}$ converges as well but with high probability instead of almost surely, since $\frac{t}{K}$ may not be bounded for all $t\in\mathbb{N}$. However, both two scalars are estimating the average trajectory length $\mathbb{E}_\mu[T]$ and are close.

The update rule is

$$\eta_{t+1} = \eta_t + \alpha_t\lambda_2(H(1-\gamma)\phi(s_t)^\top\theta_t - 1 - \eta_t). \tag{14}$$

$$\theta_{t+1} = \theta_t - \alpha_t(\phi(s_t)(\phi(s_t)^\top\theta_t - y_t) + \lambda_2\eta_t H(1-\gamma)\phi(s_t) + \lambda_1\theta_t), \tag{15}$$

where α_t is the learning rate. We combine the system of equations into

$$d_{t+1} = d_t + \alpha_t(G_{t+1}d_t + g_{t+1}),$$

where $d_{t+1} = \begin{bmatrix}\theta_{t+1}\\ \eta_{t+1}\end{bmatrix}$ denotes the concatenation of parameters, and update matrices are

$$G_{t+1} = \begin{bmatrix}-\phi(s_t)\phi(s_t)^\top - \lambda_1 I & -\lambda_2 H(1-\gamma)\phi(s_t)\\ \lambda_2 H(1-\gamma)\phi(s_t) & -\lambda_2\end{bmatrix} \text{ and } g_{t+1} = \begin{bmatrix}\phi(s_t)y_t\\ -\lambda_2\end{bmatrix}.$$

Assumption 5.1. 1. Φ has linearly independent columns.

2. Each feature vector $\phi(s)$ has its L2-norm bounded by L.
3. The behaviour policy μ induces an irreducible Markov chain on $\mathcal{S}$ and moreover, for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, $\mu(a|s)>0$ if $\pi(a|s)>0$.
4. The stepsize sequence $\{\alpha_t\}$ is deterministic and eventually nonincreasing, and satisfies $\alpha_t\in(0,1]$, $\sum_t\alpha_t=\infty$, and $\sum_t\alpha_t^2<\infty$.

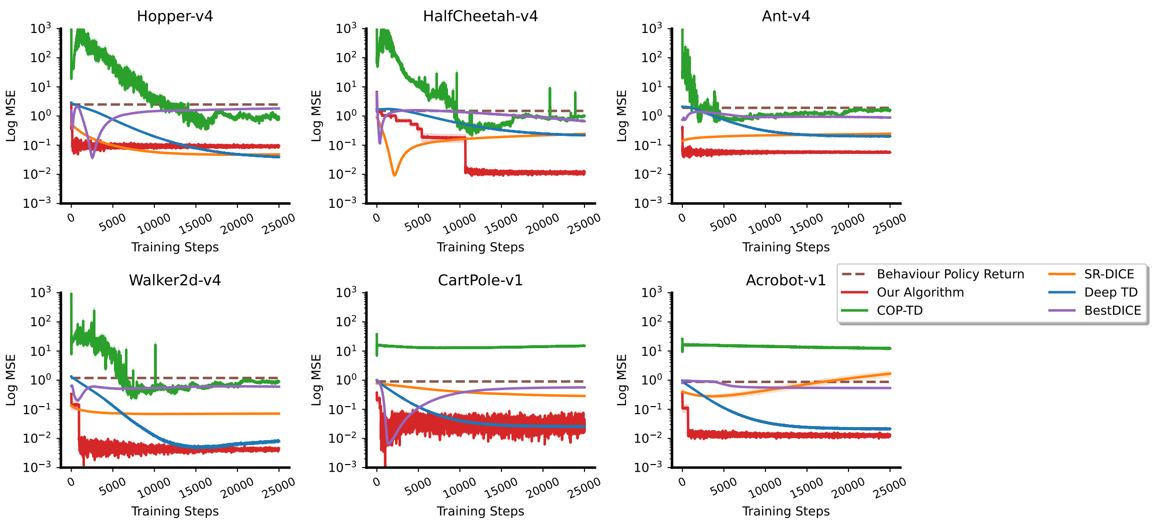


Figure 1: This figure presents the mean square error of estimating the objective $J(\pi)$ in the log scale for each task. Our method, as the red line, shows dominant behaviour on most tasks and comparable behaviour on Hopper and CartPole.

These four assumptions are also assumed for ETD convergence analysis and are common for analyzing the asymptotic behaviours of linear update rules.

Define two matrices

$$G = \begin{bmatrix} -\Phi^\top D_\mu \Phi - \lambda_1 I & -\lambda_2 H(1-\gamma)\Phi^\top d_\mu \\ \lambda_2 H(1-\gamma) d_\mu^\top \Phi & -\lambda_2 \end{bmatrix}, \text{ and } g = \begin{bmatrix} \frac{1}{(1-\gamma)\mathbb{E}_\mu[T]} \Phi^\top D_\mu y \\ -\lambda_2 \end{bmatrix},$$

where $y \in \mathbb{R}^{\mathcal{S}}$ denotes the density ratio $\frac{d_{\pi,\gamma}(s)}{d_\mu(s)}$.

Incremental updates under our losses give convergence with linear function approximation. The proof follows the convergence analysis of ETD and is presented in Appendix B. Intuitively, our correction gives a consistent estimator with variance controlled by the discount factor, and thus, the convergence follows.

Theorem 5.2. *Based on Assumption 4.2 and 5.1, we have*

$$d_t \to -G^{-1} g \text{ a.s.} \tag{16}$$

which gives the same fixed point for minimizing the mean square error to the true density ratio, which is $\mathbb{E}_{S_t \sim \mathcal{D}}\left[\frac{1}{2}\left(f_{\theta_{mse}}(S_t) - \frac{1}{(1-\gamma)\mathbb{E}_\mu[T]}\frac{d_{\pi,\gamma}(S_t)}{d_\mu(S_t)}\right)^2\right]$ *with the same regularizations.*

6 Experiments

We perform OPE on classic control and MuJoCo (Todorov et al. 2012) tasks to evaluate our method and compare it with other distribution correctors, including COP-TD, BestDice, and SR-DICE. Additionally, we include two simple baselines: the average reward, which represents the objective under the behaviour policy, and off-policy TD. In these OPE tasks, the target policy is trained using PPO (Schulman et al. 2017), which can achieve high-performing policies; for example, the agent for CartPole receives above 410 return, close to the optimal return of 500. For discrete actions, the behaviour policy is a combination of the target policy and the uniform random policy. For continuous actions, the behaviour policy is obtained by increasing the variance of the Gaussian target policy.

The hyperparameters are tuned using a dataset with 4000 transitions coming from trajectories each of length 100. The discount factor is fixed at 0.95, which is a common choice. The random policy weights 0.3 in the behaviour policy for discrete-action tasks, and the variance is doubled for continuous-action tasks. We selected the combination of hyperparameters that yields the lowest objective estimation error, averaged across all tasks. The results are averaged among 10 seeds, and the variance is tiny due to similar rewards received per step for each run.

Our results in Figure 1 show that most of the existing methods underperform compared to off-policy TD, confirming prior work (Fujimoto et al. 2021). Only SR-DICE, in the orange line, can give comparable behaviour. Our method, as the red line, shows dominant behaviour on most tasks and comparable behaviour on Hopper and CartPole. More surprisingly, it gives the fastest convergence to a stable low error. Also, in Figure 1, our method is tuned specifically to the trajectory length, the randomness of the behaviour policy, the size of the dataset, and the discount factor, which gives a small advantage to our algorithm. Thus, in the next step, we test out the robustness against more settings, as illustrated in Figure 2. This figure presents the results of the Walker task, while results for other tasks are provided in Appendix C.

The top-left subfigure of Figure 2 examines robustness against different discount factors. Our algorithm proves less robust to changes in the discount factor and loses its leading performance, yielding worse results than TD and SR-DICE. Because altering the discount factor significantly changes the regression targets for the entire dataset, it is reasonable that our method would require re-tuning for each discount factor to achieve optimal performance.

When the discount factor is fixed at 0.95, our method generally maintains a dominant or at least comparable performance relative to other baselines, except in Acrobot and Hopper. In Acrobot, it still outperforms other density-ratio estimators in most settings and is on par with TD; in Hopper, both our method and SR-DICE perform similarly to TD. Also, an ablation study without the distribution regularization is given in Appendix C.

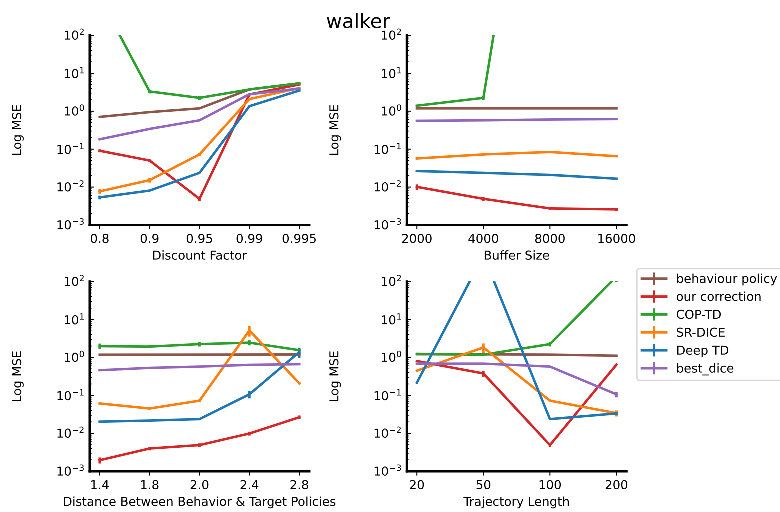


Figure 2: We evaluate the robustness of our method under varying dataset sizes, trajectory lengths, behaviour policies, and discount factors. This figure presents the results of the Walker task. Our algorithm is less robust only to changes in the discount factor and excessively long trajectory length.

7 Conclusion

We introduced AVG-DICE, a novel regression-based estimator for the stationary state density ratio. Our key contributions include deriving an alternative form of the state density ratio, proposing a novel distribution corrector and designing the learning algorithm for our distribution corrector. Furthermore, we showed that incremental updates converge under linear function approximation, demonstrating that the resulting fixed point coincides with the minimum MSE solution to the true ratio up to regularization. Empirical results on discrete and continuous tasks confirmed that AVG-DICE provides stable and accurate off-policy evaluation. Looking forward, integrating this density ratio correction into policy gradient algorithms could address distribution mismatches more effectively than current conservative policy updates.

References

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pp. 30–37. Elsevier, 1995.

Fengdi Che, Gautham Vasan, and A Rupam Mahmood. Correcting discount-factor mismatch in on-policy policy gradient methods. In *International Conference on Machine Learning*, pp. 4218–4240. PMLR, 2023.

Fengdi Che, Chenjun Xiao, Jincheng Mei, Bo Dai, Ramki Gummadi, Oscar A Ramirez, Christopher K Harris, A Rupam Mahmood, and Dale Schuurmans. Target networks and over-parameterization stabilize off-policy bootstrapping with function approximation. *arXiv preprint arXiv:2405.21043*, 2024.

Scott Fujimoto, David Meger, and Doina Precup. A deep reinforcement learning approach to marginalized importance sampling with the successor representation. In *International Conference on Machine Learning*, pp. 3518–3529. PMLR, 2021.

Carles Gelada and Marc G Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3647–3655, 2019.

Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford university press, 2020.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *International Conference on Machine Learning*, pp. 1372–1383. PMLR, 2017.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems*, 32, 2019.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in neural information processing systems*, 33:16666–16676, 2020.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Doina Precup, Richard S Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In *ICML*, pp. 417–424, 2001.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, A Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*, 17(73):1–29, 2016.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

JN Tsitsiklis and B Van Roy. An analysis of temporal-difference learning with function approximationtechnical. *Rep. LIDS-P-2322). Lab. Inf. Decis. Syst. Massachusetts Inst. Technol. Tech. Rep*, 1996.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR, 2020.

Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 33:6551–6561, 2020.

Huizhen Yu. On convergence of emphatic temporal-difference learning. In *Conference on learning theory*, pp. 1724–1751. PMLR, 2015.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2020a.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, pp. 11194–11203. PMLR, 2020b.

Supplementary Materials

The following content was not necessarily subject to peer review.

Appendix A: Derivation of Our Distribution Corrector

Proof of the Consistency Theorem

We fist present a theorem used for the proof.

Theorem 7.1 (Theorem 1.0.2 Ergodic theorem, Norris (1998))**.** *Let $\mathcal{M}$ be an irreducible and positive recurrent Markov decision process for all policies. Then, for each state s,*

$$\mathbb{P}\left(\frac{\sum_{t=0}^{T-1}\mathbb{1}[S_t = s]}{T} \rightarrow \frac{1}{\mathbb{E}[\tau_s^+(s)]} \text{ as } T \rightarrow \infty\right) = 1.$$

Next, we present our main theorem and show that our correction term equals the distribution ratio, $\frac{d_{\pi,\gamma}(s)}{d_\mu(s)}$ and thus, this term successfully corrects the state distribution shift.

Theorem 7.2 (Consistency)**.** *Given*

- *a finite Markov decision process,*
- *a dataset $\mathcal{D}$ collected under a behaviour policy μ, and*
- *a target policy π*

such that Assumption 4.1 is satisfied, then for state s with

$$d_{\pi,\gamma}(s) > 0,$$

we have the density ratio equal

$$\frac{d_{\pi,\gamma}(s)}{d_\mu(s)} = \lim_{n\rightarrow\infty}\frac{n}{K}(1-\gamma)\mathbb{E}_{t\sim I_s}[\gamma^{time_t}\rho_{prod,t}]. \tag{17}$$

Proof. **Reformulate the RHS.**

When sampling $t \sim I_s$ uniformly, the probability equals $\dfrac{\mathbb{1}[S_t = s]}{\sum_{k=1}^{n}\mathbb{1}[S_k = s]}$. Furthermore,

$$\frac{n}{K}(1-\gamma)\mathbb{E}_{t\sim I_s}[\gamma^{\text{time}_t}\rho_{\text{prod},t}] = \frac{n}{K}(1-\gamma)\sum_{t=1}^{n}\frac{\mathbb{1}[S_t = s]\gamma^{\text{time}_t}\rho_{\text{prod},t}}{\sum_{k=1}^{n}\mathbb{1}[S_k = s]} \tag{18}$$

$$= (1-\gamma)\frac{n}{\sum_{t=0}^{n-1}\mathbb{1}[S_t = s]}\frac{1}{K}\sum_{i=1}^{K}\sum_{j\geq 0}\mathbb{1}[S_j^i = s]\gamma^j\rho_{0:j-1}^i. \tag{19}$$

Note that n is the number of transitions, and K is the number of trajectories.

Define two functions:

$$g_n(s) = \frac{n}{\sum_{t=0}^{n-1}\mathbb{1}[S_t = s]}. \tag{20}$$

$$f_n(s) = (1-\gamma)\frac{1}{K}\sum_{i=1}^{K}\sum_{j\geq 0}\mathbb{1}[S_j = s]\gamma^j\rho_{0:j-1}. \tag{21}$$

Note that

$$\frac{n}{K}(1-\gamma)\mathbb{E}_{t\sim I_s}[\gamma^{\text{time}_t}\rho_{\text{prod},t}] = g_n(s)f_n(s). \tag{22}$$

Prove the irreducibility. When studying states with non-negative discounted stationary distribution values under the target policy π, they can form an irreducible set with restarts.

$d_{\pi,\gamma}(s) > 0$ implies that $d_\mu(s) > 0$. Thus,

$$\mathbb{P}_\mu(S_j = s, \text{for some } j > 0 \text{ and } j < T) > 0.$$

Meanwhile, we have $\mathbb{P}_\mu(S_j = \varsigma, \text{for some } j > 0 | S_0 = s) > 0$ for episodic tasks. Thus, with restarts, given any two states s and s' with positive stationary distribution values,

$$\begin{aligned}
&\mathbb{P}_\mu(S_j = s' | S_0 = s) \\
&> \mathbb{P}_\mu(S_j = \varsigma, \text{for some } j > 0 | S_0 = s)\mathbb{P}_\mu(S_j = s', \text{for some } j > 0 \text{ and } j < T) \\
&> 0.
\end{aligned}$$

Prove the positive recurrence. Note that a finite and irreducible Markov chain is positive recurrent.

Prove the infinite number of trajectories. By Assumption 4.2, the termination state is positive recurrent and is visited infinitely many times as the step n goes to zero. Thus, there are infinitely many trajectories.

Compute the almost sure limit of two functions. The function $g_n(s)$ is proven to converge to $g(s) = \mathbb{E}_\mu[\tau_s^+(s)]$ by the ergodic theorem.

Apparently, $\lim_{n\to\infty} f_n(s) = \lim_{K\to\infty}(1-\gamma)\frac{1}{K}\sum_{i=1}^{K}\sum_{j\geq 0}\mathbb{1}[S_j = s]\gamma^j\rho_{0:j-1}$. By the central limit theorem, we have

$$\lim_{n\to\infty} f_n(s) = \mathbb{E}_\mu\left[(1-\gamma)\sum_{j\geq 0}\mathbb{1}[S_j = s]\gamma^j\rho_{0:j-1}\right] = d_{\pi,\gamma}(s). \tag{23}$$

Thus,

$$\begin{aligned}
L.H.S &= \lim_{n\to\infty} g_n(s) f_n(s) && (24)\\
&= g(s) f(s) && (25)\\
&= \mathbb{E}_\mu[\tau_s^+(s)] d_{\pi,\gamma}(s) && (26)\\
&= \frac{d_{\pi,\gamma}(s)}{d_\mu(s)}. && (27)
\end{aligned}$$

The last line follows Lemma 2.1, and the proof is completed. □

Proof of the Unbiasedness

Theorem 7.3 (Unbiasedness)**.** *Given*

- *a finite Markov decision process,*
- *a dataset $\mathcal{D}$ collected under a behaviour policy μ, and*
- *a target policy π*

such that Assumption 4.1 is satisfied, reweighting data by our average correction gives unbiased estimation for any function $f : \mathcal{S} \to \mathbb{R}$, that is,

$$\mathbb{E}_\mathcal{D}\left[\mathbb{E}_{S\sim\mathcal{D}}\left[c_\mathcal{D}(S) f(S)\right]\right] = \mathbb{E}_{S\sim d_{\pi,\gamma}}[f(S)], \tag{28}$$

where $\mathbb{E}_\mathcal{D}$ means expectation over trajectories sampled under the behaviour policy, and $\mathbb{E}_{S\sim\mathcal{D}}$ representing sampling states uniformly from the dataset.

Proof. Denote the sampling distribution from the dataset as $\hat{d}(s) = \frac{\sum_{t=0}^{n-1}\mathbb{1}[S_t = s]}{n}$.

As proven in Corollary 4.3 in Equation 22,

$$c_\mathcal{D}(S) = \frac{1}{\hat{d}(s)}(1-\gamma)\frac{1}{K}\sum_{i=1}^{K}\sum_{j\geq 0}\mathbb{1}[S_j^i = s]\gamma^j\rho_{0:j-1}^i. \tag{29}$$

Thus, $\mathbb{E}_{S\sim\mathcal{D}}\left[c_\mathcal{D}(S) f(S)\right] = \sum_{s\in\mathcal{S}} f(s)(1-\gamma)\frac{1}{K}\sum_{i=1}^{K}\sum_{j\geq 0}\mathbb{1}[S_j^i = s]\gamma^j\rho_{0:j-1}^i$.

After taking expectation over all K trajectories, we have

$$\mathbb{E}_{\mathcal{D}}\left[\sum_{s\in\mathcal{S}} f(s)(1-\gamma)\frac{1}{K}\sum_{i=1}^{K}\sum_{j\geq 0}\mathbb{1}[S_j^i = s]\gamma^j\rho_{0:j-1}^i\right] \tag{30}$$

$$= \sum_{s\in\mathcal{S}} f(s)(1-\gamma)\sum_{j\geq 0}\gamma^j\mathbb{P}_\pi(S_j = s) \tag{31}$$

$$= \sum_{s\in\mathcal{S}} f(s)d_{\pi,\gamma}(s). \tag{32}$$

□

Appendix B: Asymptotic Convergence

We first introduce and prove the necessary lemmas. The proof of the convergence theorem is given in the second subsection.

Proof of the Required Lemma

Denote the number of trajectories until step t by $K(t)$. Recall our update rule is $d_{t+1} = d_t + \alpha_t(G_{t+1}d_t + g_{t+1})$ where $d_{t+1} = \begin{bmatrix}\theta_{t+1}\\ \eta_{t+1}\end{bmatrix}$, $G_{t+1} = \begin{bmatrix}-\phi(s_t)\phi(s_t)^\top - \lambda_1 I & -\lambda_2 H(1-\gamma)\phi(s_t)\\ \lambda_2 H(1-\gamma)\phi(s_t) & -\lambda_2\end{bmatrix}$, and $g_{t+1} = \begin{bmatrix}\phi(s_t)y_t\\ -\lambda_2\end{bmatrix}$.

Lemma 7.4. *Define two matrices* $G = \begin{bmatrix}-\Phi^\top D_\mu\Phi - \lambda_1 I & -\lambda_2 H(1-\gamma)\Phi^\top d_\mu\\ \lambda_2 H(1-\gamma)d_\mu^\top\Phi & -\lambda_2\end{bmatrix}$ *and* $g = \begin{bmatrix}\frac{1}{(1-\gamma)\mathbb{E}_\mu[T]}\Phi^\top D_\mu y\\ -\lambda_2\end{bmatrix}$.

When Assumption 4.2 and 5.1 are satisfied, we have

1. $\frac{1}{t+1}\sum_{k=0}^{t+1} G_k \to G$ *a.s. , and* $\frac{1}{t+1}\sum_{k=0}^{t+1} g_k \to g$ *a.s. and in L1, as* $t\to\infty$.
2. *The real parts of all eigenvalues of G are strictly negative.*

Proof. **Let's prove the first point about almost sure convergence.** We will prove the convergence of each sub-matrix separately.

Note that the ergodic theorem gives that the convergence of the top-left sub-matrix of G_t as

$$\begin{aligned}&\frac{1}{t+1}\sum_{k=0}^{t+1} -\phi(s_k)\phi(s_k)^\top - \lambda_1 I\\ &\xrightarrow{t\to\infty}\sum_{s\in\mathcal{S}} -d_\mu(s)\phi(s)\phi(s)^\top - \lambda_1 I\\ &= -\Phi^\top D_\mu\Phi - \lambda_1 I.\end{aligned}$$

Similar convergence can be gained for the term $\lambda_2 H(1-\gamma)\phi(s_t)$ by the ergodic theorem as well.

Combining these two results, we gain the almost sure convergence of G_t.

Now we analyze the term $y_t\phi(s_t) = \sum_s \phi(s)\sum_{i\in I_s(t)} y_i$ where $I_s(t)$ denotes the showing up steps for a state a until step t. It can be further expressed as

$$\begin{aligned}
&\frac{1}{t+1}\sum_{k=0}^{t} y_t\phi(s_t)\\
&=\sum_s \frac{K(t)}{t+1}\phi(s)\frac{1}{1-\gamma}(1-\gamma)\frac{1}{K(t)}\sum_{i\in I_s(t)} y_i\\
&\to \frac{1}{1-\gamma}\sum_s d_{\pi,\gamma}(s)\phi(s)\frac{1}{\mathbb{E}_\mu[T]}\\
&=\frac{1}{(1-\gamma)\mathbb{E}_\mu[T]}\sum_s d_\mu(s)\frac{d_{\pi,\gamma}(s)}{d_\mu(s)}\phi(s).
\end{aligned}$$

Note that the third line uses the convergence of $\frac{K(t)}{t+1}$ to $\frac{1}{\mathbb{E}_\mu[T]}$. $K(t)$ goes to infinity as $t\to\infty$ since the recurrence time for the termination state has a finite expectation by Assumption 4.2. Similarly, the ergodic theorem implies the almost sure convergence of $\frac{t+1}{K(t)}$ to $\mathbb{E}_\mu[T]$.

Furthermore, by central limit theorem, $(1-\gamma)\frac{1}{K(t)}\sum_{i\in I_s(t)} y_i$ converges to $d_{\pi,\gamma}(s)$. Since the limits of these two terms ar bounded, the limit of the product converges to the product of limits.

For the L1-convergence, we analyze the expectation of $y_t\phi(s_t)$. Define $T_{K(t)}$ as the termination step of $K(t)$-th trajectory.

$$\frac{1}{t+1}\sum_{k=0}^{t+1}\mathbb{E}_\mu[y_t\phi(s_t)] \tag{33}$$

$$=\frac{1}{t+1}\sum_{k=0}^{t+1}\gamma^t\mathbb{P}_\pi(S_t=s)\phi(s) \tag{34}$$

$$=\sum_s \frac{K(t)}{t+1}\phi(s)\frac{1}{1-\gamma}(1-\gamma)\frac{1}{K(t)}[K(t)\sum_{j\geq 0}\gamma^j\mathbb{P}_\pi(S_j=s)+\sum_{j=0}^{t-T_{K(t)}}\gamma^j\mathbb{P}_\pi(S_j=s)] \tag{35}$$

$$\to \frac{1}{(1-\gamma)\mathbb{E}_\mu[T]}\sum_s d_\mu(s)\frac{d_{\pi,\gamma}(s)}{d_\mu(s)}\phi(s). \tag{36}$$

Note that $\frac{1}{K(t)}\sum_{j=0}^{t-T_{K(t)}}\gamma^j\mathbb{P}_\pi(S_j=s)\to 0$ as $K(t)\to\infty$ and $t\to\infty$.

Let's prove the second point about eigenvalues.

Let $\vartheta\in\mathbb{C}$, $\vartheta\neq 0$ be a nonzero eigenvalue of G with normalized eigenvector x, that is $x^*x=1$, where x^* is the complex conjugate of x. Hence, $x^*Gx=\vartheta$, $x\neq 0$. Let $x^\top=(x_1^\top, x_2)$, where $x_1\in\mathbb{C}^d$ and $x_2\in\mathbb{C}$. We can verify that

$$\vartheta = -x_1^*(\Phi^\top D_\mu\Phi+\lambda_1 I)x_1+\lambda_2 x_2^* H(1-\gamma)d_\mu^\top\Phi x_1-\lambda_2 x_1^* H(1-\gamma)d_\mu\Phi^\top x_2-\lambda_2 x_2^* x_2. \tag{37}$$

Since $d_\mu^\top\Phi$ is real, $\lambda_2 x_1^* H(1-\gamma)d_\mu\Phi^\top x_2=(\lambda_2 x_2^* H(1-\gamma)d_\mu^\top\Phi x_1)^*$. It yields that the real part of their difference equals zero. Therefore, we have the real part of ϑ, denoted by $\mathcal{R}(\vartheta)$, equals

$$\mathcal{R}(\vartheta)=-x_1^*(\Phi^\top D_\mu\Phi+\lambda_1 I)x_1-\lambda_2 x_2^* x_2. \tag{38}$$

By the first point in Assumption 5.1, we have $-x_1^*(\Phi^\top D_\mu\Phi+\lambda_1 I)x_1\geq 0$, where the equality holds iff $x_1=0$. At least one of $\{x_1,x_2\}$ is nonzero. Consequently, we have $\mathcal{R}(\vartheta)<0$. ☐

Proof of the Convergence

Theorem 7.5. *Based on Assumption 4.2 and 5.1, and Lemma 7.4, we have*

$$d_t\to -G^{-1}g \text{ a.s.} \tag{39}$$

which is the same fixed point for minimizing the following mean square error loss in Equation 40.

The loss given a dataset $\mathcal{D}$ to the true density ratio is written as

$$\min_{\theta_{\text{mse}}} \mathcal{L}(\theta_{\text{mse}};\mathcal{D}) := \mathbb{E}_{s_t\sim\mathcal{D}}\left[\left(f_{\theta_{\text{mse}}}(s_t) - \frac{1}{(1-\gamma)\mathbb{E}_\mu[T]}\frac{d_{\pi,\gamma}(s_t)}{d_\mu(s_t)}\right)^2\right] + \frac{\lambda_1\|\theta_{\text{mse}}\|_2^2}{2} + \lambda_2\left(max_{\eta\in\mathbb{R}}\mathbb{E}_{s_t\sim\mathcal{D}}[\eta\frac{n}{K}(1-\gamma)f_{\theta_{\text{mse}}}(s_t)-\eta] - \frac{\eta^2}{2}\right). \tag{40}$$

Proof. First, we can verify some properties on our labels. Based on these properties and L1-convergence of $\frac{1}{t+1}\sum_{k=0}^t g_k$, we conclude that our label $(y_t, \phi(s_t)y_t)$ gives a unique invariant probability and is ergodic.

The proof is the same as the corresponding proofs of (Yu, 2012, Theorem 3.2 and Prop. 3.2) for the case of off-policy LSTD.

1. For any initial value of $\rho_{0:-1}$, $\sup_{t\geq 0}\mathbb{E}[|(y_t,\phi(s_t)y_t)\|] < \infty$.
2. Let $(y_t,\phi(s_t)y_t)$ and $(\hat{y}_t,\phi(s_t)\hat{y}_t)$ be defined by the same recursion and the same random variables, but with different initial conditions $\rho_{0:-1} \neq \hat{\rho}_{0:-1}$. Then, $y_t - \hat{y}_t \to 0$ a.s. and $\phi(s_t)y_t - \phi(s_t)\hat{y}_t \to 0$ a.s..
3. $Z_t = (S_t, A_t, y_t, \phi(s_t)y_t)$ is a weak Feller Markov chain and bounded in probability.

The proof follows ETD, since $y_t = \gamma^{\text{time}_t}\rho_{\text{prod},t}$ is a term in the ETD traces. For the second term, the difference between traces with different initializations for our correction and ETD is the same, so their proof also works here. The proof for the third claim follows the ETD paper.

Three conditions are required to use Theorem 6.1.1 in Kushner and Yin (2003) and follow the ETD proof (Theorem 4.1). Define $\xi_t = (y_t, S_t, A_t, S_{t+1})$ and $h(d,\xi_t) = G_t d + g_t$.

1.

$$\frac{1}{t+1}\sum_{k=0}^t G_k \to G \text{ and } \frac{1}{t+1}\sum_{k=0}^t g_k \to g \text{ almost surely.} \tag{41}$$

2. There exist nonnegative measurable functions $g_1(d)$, $g_2(\xi)$ such that $\|h(d,\xi)\| \leq g_1(d)g_2(\xi)$ such that $g_1(d)$ is bounded on each bounded set, $\sum_{t\geq 0}\mathbb{E}[g_2(\xi)] < \infty$, and $\frac{1}{t+1}\sum_{k=0}^t (g_2(\xi_k) - \mathbb{E}[g_2(\xi_k)]) \to 0$ almost surely.
3. There exist nonnegative measurable functions $g_3(d)$, $g_4(\xi)$ such that for each d and d', $\|h(d,\xi) - h(d',\xi)\| \leq g_3(d-d')g_4(\xi)$ such that $g_3(d)$ is bounded on each bounded set, $g_3(d)\to 0$ as $d\to 0$, $\sum_{t\geq 0}\mathbb{E}[g_4(\xi)] < \infty$, and $\frac{1}{t+1}\sum_{k=0}^t (g_4(\xi_k) - \mathbb{E}[g_4(\xi_k)]) \to 0$ almost surely.

In our proof, the function $h(d,\xi)$ equals

$$h(d,\xi) = \begin{bmatrix} -\phi(s)\phi(s)^\top - \lambda_1 I & -\lambda_2 H(1-\gamma)\phi(s_t) \\ \lambda_2 H(1-\gamma)\phi(s_t) & -\lambda_2 \end{bmatrix} d + \begin{bmatrix}\phi(s)y \\ -\lambda_2\end{bmatrix}. \tag{42}$$

Then, for the second and third points, we first bound the norm of the matrix as followings.

$$\|\begin{bmatrix} -\phi(s)\phi(s)^\top - \lambda_1 I & -\lambda_2 H(1-\gamma)\phi(s_t) \\ \lambda_2 H(1-\gamma)\phi(s_t) & -\lambda_2 \end{bmatrix} \tag{43}$$

$$\leq \|\begin{bmatrix} -\phi(s)\phi(s)^\top - \lambda_1 I & 0 \\ 0 & 0 \end{bmatrix}\| + \|\begin{bmatrix} 0 & -\lambda_2 H(1-\gamma)\phi(s_t) \\ \lambda_2 H(1-\gamma)\phi(s_t) & -\lambda_2 \end{bmatrix}\| \tag{44}$$

$$\leq \|-\phi(s)\phi(s)^\top - \lambda_1 I\| + \sqrt{(2\|\lambda_2 H(1-\gamma)\phi(s_t)\|^2)}\|+\lambda_2 \tag{45}$$

$$\leq L^2 + \lambda_1 + \sqrt{2}\lambda_2 H(1-\gamma)L + \lambda_2. \tag{46}$$

$$\|h(d,\xi)\| \leq \| \begin{bmatrix} -\phi(s)\phi(s)^\top - \lambda_1 I & -\lambda_2 H(1-\gamma)\phi(s_t) \\ \lambda_2 H(1-\gamma)\phi(s_t) & -\lambda_2 \end{bmatrix} \| \|d\| + \| \begin{bmatrix} \phi(s)y \\ -\lambda_2 \end{bmatrix} \| \tag{47}$$

$$\leq (L^2 + \lambda_1 + \sqrt{2}\lambda_2 H(1-\gamma)L + \lambda_2)\|d\| + (\|\phi(s)y\|\lambda_2) \tag{48}$$

$$\leq (L^2 + \lambda_1 + \sqrt{2}\lambda_2 H(1-\gamma)L + \lambda_2 + Ly)(\|d\| + 1). \tag{49}$$

Thus, $g_1(d) = (\|d\| + 1)$ and $g_2(\xi) = L^2 + \lambda_1 + \sqrt{2}\lambda_2 H(1-\gamma)L + \lambda_2 + Ly$.

We can bound the function norm using the matrix norm bound in Equation 46.

$$\|h(d,\xi) - h(d',\xi)\| \leq \| \begin{bmatrix} -\phi(s)\phi(s)^\top - \lambda_1 I & -\lambda_2 H(1-\gamma)\phi(s_t) \\ \lambda_2 H(1-\gamma)\phi(s_t) & -\lambda_2 \end{bmatrix} \| \|d - d'\| \tag{50}$$

$$\leq (L^2 + \lambda_1 + \sqrt{2}\lambda_2 H(1-\gamma)L + \lambda_2)\|d - d'\|. \tag{51}$$

Thus, $g_3(d) = \|d\|$ and $g_4(\xi) = L^2 + \lambda_1 + \sqrt{2}\lambda_2 H(1-\gamma)L + \lambda_2$ is a constant.

To show the fixed point is the same as minimizing the MSE to the true density ratio, we need to repeat the convergence proof for the new loss. But the only change is in the regression target and all other steps follow. □

Appendix C: Experimental Materials

The hyperparameters of COP-TD are tuned the same as our method in the setting of dataset size 4000, trajectory length 100, discount factor 0.95 and randomness coefficient 0.3 for discrete-action tasks and 2.0 for continuous-action tasks. Only one combination of the hyperparameters is used for all tasks.

For our algorithms, we test out the combination from parameter regularization coefficient $\lambda_1 \in [0, 0.001, 0.01, 0.1]$, distribution regularization parameter $\lambda_2 \in [0.5, 2, 10, 20]$, and learning rate $\alpha \in [0.00005, 0.0001, 0.0005, 0.001, 0.005]$.

The neural network is set to be a two-hidden-layer neural network with hidden units 256, which is the setting used by SR-DICE. The batch size is set the same as SR-DICE, equaling 512.

The final choice of hyperparameters is shown in Table 1.

Parameter Regularizer λ_1	0.001
Distribution Regularizer λ_2	0.5
Learning Rate	0.0005
Activation	ReLU

Table 1

Ablation Study

When training without the distribution regularization, the hyperparameters are also tuned with the regularization coefficient fixed, $\lambda_2 = 0$. Notice that the training step is much fewer than Figure 1. The results are plotted on a validation set. Without the distribution regularization, the ratio model is not learning except on CartPole.

Robustness

The top-left subfigure of Figure 2 examines robustness against different discount factors. The main message is that our algorithm proves less robust to changes in the discount factor and would require re-tuning for each discount factor to achieve optimal performance. Our algorithm is robust to other changes except for a trajectory length that is too long. Results for other tasks are presented in Figure 4 and 5 and the conclusion holds for all tasks.

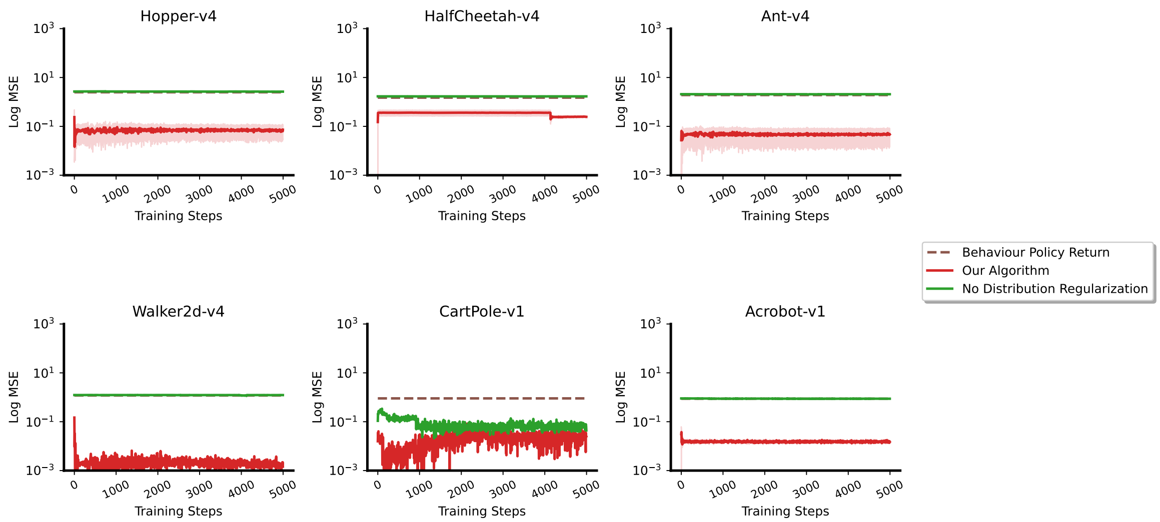
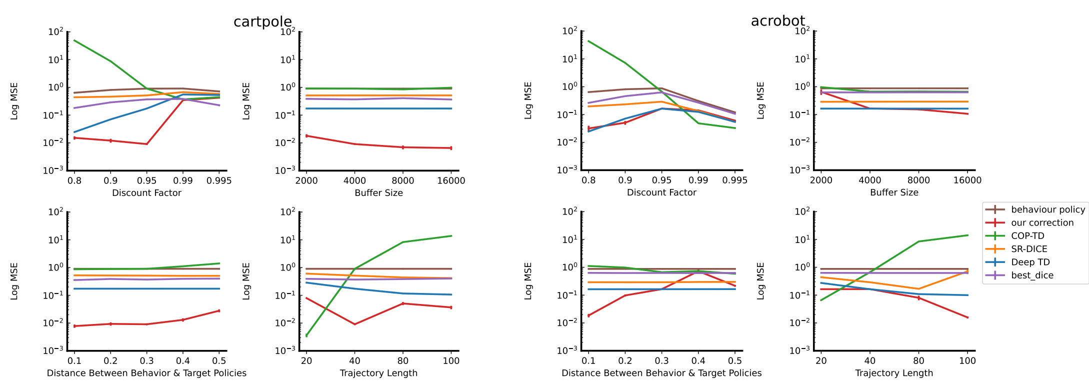


Figure 3: This figure presents the mean square error of estimating the objective $J(\pi)$ in the log scale for each task. We present our method with and without the distribution regularization.

Turning to the bottom-right subfigure of Figure 2, performance degrades when the trajectory length is set to 200. A similar phenomenon is observed in Hopper and HalfCheetah. We propose two hypotheses for this drop. First, our method may struggle with very long trajectories, suggesting that users might benefit from truncating trajectories since longer horizons lead to heavily discounted and, thus, tiny labels. Second, the total dataset size is fixed at $4,000$, so increasing the length of each trajectory decreases the number of available trajectories. The approximation error decays sublinearly with the number of trajectories; thus, fewer trajectories can hinder performance.



(a) (b)

Figure 4: This figure shows the robustness results on discrete-action tasks.

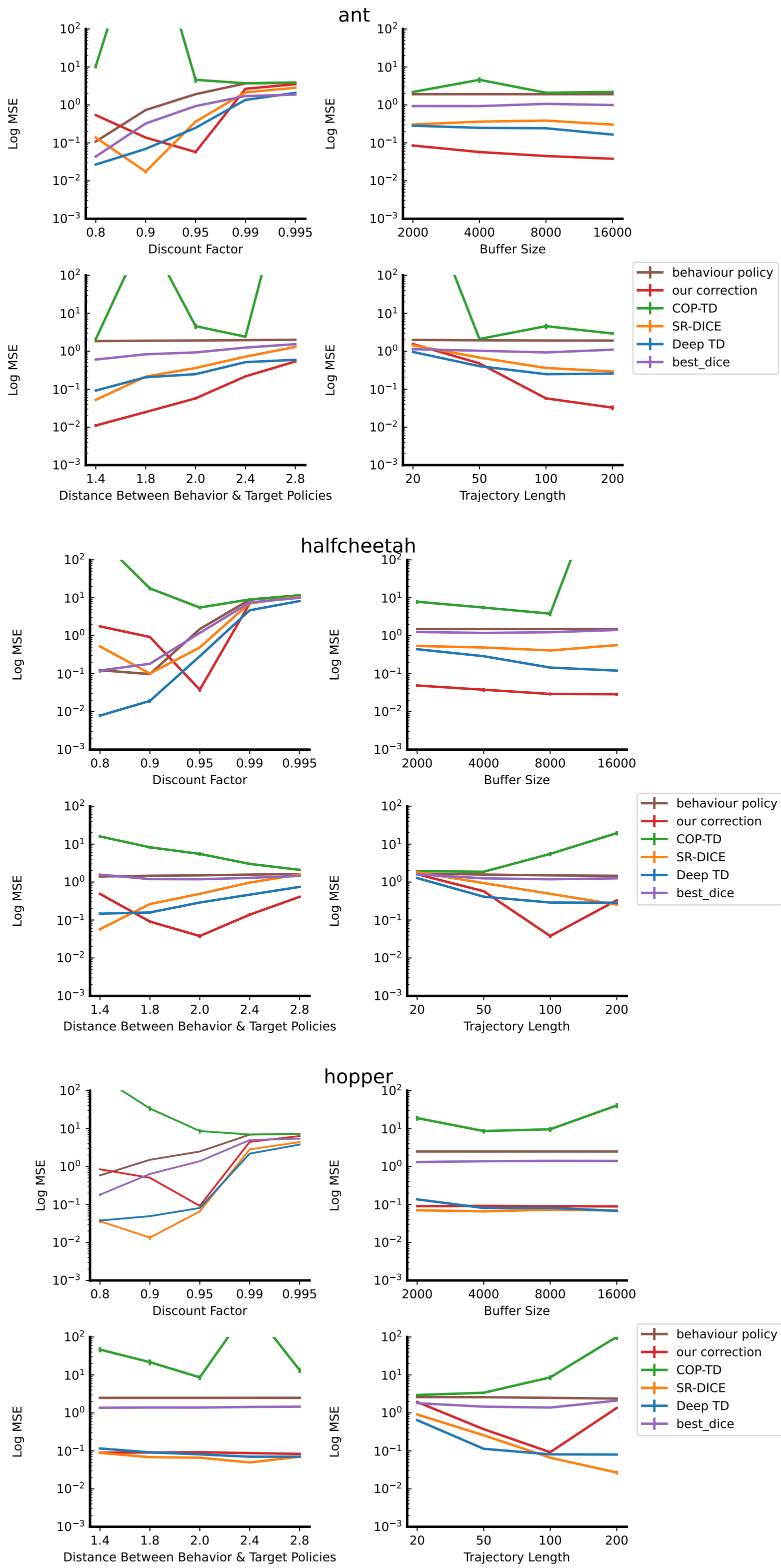


Figure 5: This figure shows the robustness results on continuous-action tasks.