Marginalized Operators for Off-policy Reinforcement Learning

Yunhao Tang
DeepMind

Mark Rowland
DeepMind

Rémi Munos
DeepMind

Michal Valko
DeepMind

Abstract

In this work, we propose marginalized operators, a new class of off-policy evaluation operators for reinforcement learning. Marginalized operators strictly generalize generic multi-step operators, such as Retrace, as special cases. Marginalized operators also suggest a form of sample-based estimates with potential variance reduction, compared to sample-based estimates of the original multi-step operators. We show that the estimates for marginalized operators can be computed in a scalable way, which also generalizes prior results on marginalized importance sampling as special cases. Finally, we empirically demonstrate that marginalized operators provide performance gains to off-policy evaluation and downstream policy optimization algorithms.

1 Introduction

In many applications of reinforcement learning (RL), it is useful to be able to learn about one policy using data generated by a different policy, such as exploratory data [1], expert data [2] or even offline data [3]; this is the problem of off-policy learning. To successfully learn in such scenarios, off-policy algorithms must be able to safely deal with discrepancies between the data-generating policy and policy of interest. As a fundamental building block of generic off-policy algorithms, off-policy evaluation studies the problem of estimating value functions of a target policy π with data collected under behavior policy μ.

A distinction is often drawn in off-policy learning between *online* and *offline* learning. In the *online* setting, where RL agents keep collecting new data, most prior work focuses on multi-step operator-based methods (e.g., [4, 5, 6, 7]). These methods equate policy evaluations to solving for fixed points of contractive operators. In this case, a central idea is *bootstrapping*, where new estimates build on old estimates in an iterative fashion. As a result of contractive operators, the sequence of output from the algorithm forms increasingly accurate predictions to the true target values. This is especially desirable in many practical online setups where the target policy might slowly change over time (e.g., policy optimization), where predictions for the new policy could extract useful information from predictions for old policies.

On the other hand, in the *offline* setting where no further data collection is possible, much work builds on importance sampling (IS) [4, 8, 9, 10, 11, 12, 13, 14, 15]. Popular approaches for variance reduction in importance sampling are based on marginalized IS [10, 14] which has also shown promises even when combined with function approximations for high-dimensional input spaces [11, 13, 16]. However, since the offline problems only require a *single* numerical prediction, most algorithms do not naturally incorporate the notion of *bootstrapping* out-of-the-box. As a result, despite some recent efforts [17], it is in general challenging to directly apply such methods to *online* off-policy learning.

Motivated by the disparity between these two lines of work, we propose marginalized operators, a new family of off-policy evaluation operators that generalize multi-step operators as special cases (Section 3). Marginalized operators suggest new stochastic estimates to the equivalent multi-step operators, with connections to marginalized IS (Section 4). Under this framework, we also consider *estimated* marginalized operators (Section 5), which can be computed with estimates in a scalable manner, and can be analyzed as estimators in their own right. Finally, we show that the new operators provide performance gains on both policy evaluation and downstream optimization (Section 6).

Our discussions are limited to multi-step operators constructed as a weighted mixture of Bellman errors across different time steps. As a result, Q^π is the unique fixed point of such operators; these exclude operators which explicitly bias the fixed point in exchange for faster contraction rate, such as the uncorrected n-step operator. See [7] for a comprehensive discussion.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

2 Background

2.1 Markov decision processes

Consider the setup of a Markov decision process (MDP) [18] with an infinite horizon. At any discrete time $t \geq 0$, the agent is in state $x_t \in \mathcal{X}$, takes an action $a_t \in \mathcal{A}$. The agent first receives an immediate random reward $r_t = r(x_t, a_t)$ with mean $\bar{r}(x_t, a_t)$, and then transitions to a next state $x_{t+1} \sim p(\cdot|x_t, a_t)$. We assume rewards are deterministic, but most results extend naturally to the stochastic case. Below, we will discuss when such extensions do not hold. Let policy $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ be a mapping from states to distributions over actions. Let $\gamma \in [0, 1)$ be a discount factor, define the Q-function $Q^\pi(x, a) \coloneqq \mathbb{E}_\pi \left[\sum_{t=0}^\infty \gamma^t r_t \mid x_0 = x, a_0 = a\right]$ and value function $V^\pi(x) \coloneqq \mathbb{E}_\pi \left[\sum_{t=0}^\infty \gamma^t r_t \mid x_0 = x\right]$. Here, $\mathbb{E}_\pi [\cdot]$ denotes that the trajectories $(x_t, a_t, r_t)_{t=0}^\infty$ are generated under policy π.

2.2 Multi-step off-policy evaluation

Consider off-policy evaluation where π is the target policy and μ is the behavior policy, where we assume $\text{supp}\,(\pi(\cdot|x)) \subset \text{supp}\,(\mu(\cdot|x)), \forall x \in \mathcal{X}$. Given a trajectory $(x_t, a_t, r_t)_{t=0}^\infty$ generated under μ and a Q-function Q, we define the TD error at time t as $\Delta_t^\pi Q \coloneqq \bar{r}_t + \gamma \mathbb{E}_{x' \sim p(\cdot|x_t, a_t)} [Q(x', \pi(x'))] - Q(x_t, a_t)$. Here, we adopt the notation $Q(x, \pi(x)) \coloneqq \mathbb{E}_{a \sim \pi(\cdot|x)} [Q(x, a)]$. The multi-step off-policy evaluation operators $\mathcal{R}^c$ [6] define the step-wise trace coefficient $c_t \in \mathbb{R}$ per time step t, where in general $c_t = c(\{x_s, a_s\}_{s \leq t})$ is a function of the of the past $(x_s, a_s)_{s \leq t}$. The Q-function estimate $\mathcal{R}^c Q(x, a)$ at the starting pair (x, a) is computed as

$$Q(x, a) + \mathbb{E}_\mu \left[\sum_{t \geq 0} \gamma^t (\Pi_{1 \leq s \leq t} c_s) \Delta_t^\pi Q \,\middle|\, x_0 = x, a_0 = a \right], \tag{1}$$

where we define $(\Pi_{1 \leq s \leq t} c_s) = 1$ when $t = 0$. When $0 \leq c_t \leq \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}$, it can be shown that Q^π is the unique fixed point to $\mathcal{R}^c Q = Q$ [6]. As an important example, let $c_t = \mathbb{I}[t \leq 0]$, the operator $\mathcal{R}^c$ reduces to the one-step Bellman operator $\mathcal{T}^\pi Q(x, a) \coloneqq r_0 + \gamma \mathbb{E}_\pi [Q(x_1, \cdot)]$. In this case, the traces c_t are *cut off* beyond the first time step, which prevents the algorithm from bootstrapping from the rest of the trajectory. In many cases, the coefficient $c_t = c(x_t, a_t)$ is Markovian if it only depends on (x_t, a_t). Notable examples include importance sampling $c_t = \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}$, Retrace $c_t = \lambda \min\{\bar{c}, \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}\}$ [6], tree backup $c_t = \pi(a_t|x_t)$ [4] and Q$^\pi(\lambda)$ $c_t = \lambda$ [5].

2.3 Off-policy evaluation via marginalized importance sampling

We start by introducing the discounted visitation distribution $d_{x,a}^\pi(x', a') \coloneqq (1 - \gamma) \sum_{t \geq 0} \gamma^t \mathbb{P}_\pi(x_t = x', a_t = a'|x_0 = x, a_0 = a)$ where (x, a) are the starting state-action pair. The discounted visitation distribution $d_{x,a}^\pi(x', a')$ and value functions $Q^\pi(x, a)$ are related as follows [18],

$$Q^\pi(x, a) = (1 - \gamma)^{-1} \mathbb{E}_{(x', a') \sim d_{x,a}^\pi} [r(x', a')] \,. \tag{2}$$

Assume the off-policy data is sampled under $d_{x,a}^\mu(x', a')$. Let $w_{x,a}^{\pi,\mu}(x', a') \coloneqq \frac{d_{x,a}^\pi(x', a')}{d_{x,a}^\mu(x', a')}$. One could express $Q^\pi(x, a)$ via marginalized IS [14, 10],

$$Q^\pi(x, a) = (1 - \gamma)^{-1} \mathbb{E}_{(x', a') \sim d_{x,a}^\mu} [w(x', a') r(x', a')] \,.$$

For convenience, let $w^{\pi,\mu} \in \mathbb{R}^{(\mathcal{X} \times \mathcal{A}) \times (\mathcal{X} \times \mathcal{A})}$ be a matrix such that $w_{x,a}^{\pi,\mu}(x', a')$ is the entry at (x, a, x', a'). Since marginalized IS ratios are generally unknown, it is necessary to construct estimates $w_\psi \approx w^{\pi,\mu}$. There are a number of algorithms which carry out the estimation in a scalable way, which we will detail in Section 5.

Remarks on notations. Note that $Q^\pi : \mathbb{R}^{\mathcal{X} \times \mathcal{A}} \mapsto \mathbb{R}$ ($w^{\pi,\mu} : \mathbb{R}^{(\mathcal{X} \times \mathcal{A}) \times (\mathcal{X} \times \mathcal{A})} \mapsto \mathbb{R}$) are by defintion functions. To facilitate derivations, we abuse notations and also treat them as vectors (matrices) such that $Q^\pi \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}|}$($w^{\pi,\mu} \in \mathbb{R}^{|\mathcal{X}||\mathcal{A}| \times |\mathcal{X}||\mathcal{A}|}$). As such, $Q^\pi(x, a)$ can be both interpreted function evaluation and vector indexing at (x, a).

3 Marginalized Off-Policy Evaluation Operators

The marginalized off-policy evaluation operator $\mathcal{M}^w : \mathbb{R}^{\mathcal{X} \times \mathcal{A}} \to \mathbb{R}^{\mathcal{X} \times \mathcal{A}}$ is defined such that its component at (x, a) is evaluated as

$$Q(x, a) + (1 - \gamma)^{-1} \mathbb{E}_{(x', a') \sim d_{x,a}^\mu} [w_{x,a}(x', a') \Delta^\pi(x', a')] \,, \tag{3}$$

where $w_{x,a}(x', a')$ are called *TD weights*. Define $\Delta^\pi(x, a) \coloneqq \bar{r}(x, a) + \gamma \mathbb{E}_{x' \sim p(\cdot|x,a)} [Q(x', \pi(x')] - Q(x, a)$ as (x, a)-dependent Bellman errors. Note the difference between $\mathbb{E}_\mu [\cdot]$ in Eqn (1), which is an expectation over trajectories $(x_t, a_t, r_t)_{t=0}^\infty$ under μ; and $\mathbb{E}_{d_{x,a}^\mu} [\cdot]$ in Eqn (3), which is an expectation under the discounted distribution.

Below, we will first characterize important properties of the marginalized operator. Then, we will show that the space of contractive marginalized operators contains the space of contractive multi-step operators.

3.1 Properties of the marginalized operator

The following proposition summarizes a few important properties of the marginalized operators.

Proposition 3.1. For any TD weights w, the Q-function Q^π is a solution to the fixed point equation $\mathcal{M}^w Q = Q$. For any $Q_1, Q_2 \in \mathbb{R}^{\mathcal{X} \times \mathcal{A}}$,

$$|\mathcal{M}^w Q_1(x, a) - \mathcal{M}^w Q_2(x, a)| \leq \eta_{x,a}^w \left\| Q_1 - Q_2 \right\|_\infty .$$

Let $\delta_{x,a} \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ be the one-hot encoding of (x,a) and let $d_{x,a}^w \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ such that $d_{x,a}^w(x',y') = w_{x,a}(x',y')d_{x,a}^\mu(x',a')$. Then define the *residual error vector*

$$E_{x,a}^w = (1-\gamma)\delta_{x,a} + \gamma(P^\pi)^T d_{x,a}^w - d_{x,a}^w,$$

which characterizes how $d_{x,a}^w$ satisfies the *balance equations*

$$(1-\gamma)\delta_{x,a} + \gamma(P^\pi)^T d - d = 0. \tag{4}$$

The local contraction rate is expressed as

$$\eta_{x,a}^w = (1-\gamma)^{-1}\left\|E_{x,a}^w\right\|_1. \tag{5}$$

The above implies that the operator is contractive when $\max_{x,a}\left\|E_{x,a}^w\right\|_1 < 1-\gamma$.

Proposition 3.1 shows that the local contraction rate $\eta_{x,y}^w$ is proportional to the L^1 norm of the residual error vector of $d_{x,y}^w$ when plugged into the balance equation. This means that in order for $\mathcal{M}^w$ to be contractive, we seek w such that it approximately satisfies the balance equation and the residual error vector is small.

Similar to the notation of $w^{\pi,\mu}$, we denote w as the matrix of TD weights. Though it is not straightforward to analytically characterize the set of w such that $\mathcal{M}^w$ is contractive, we shed light on properties of such w with some examples.

Marginalized IS ratios as a special case. The discounted visitation distribution $d_{x,a}^\pi$ is the only solution that satisfies the balance equation. When $w_{x,a} = w_{x,a}^{\pi,\mu}$, since balance equations are satisfied exactly, $\eta_{w_{x,a}^{\pi,\mu}} = 0$ and the contraction is instant $\mathcal{M}^{w^{\pi,\mu}}Q = Q^\pi, \forall Q$. Instead of requiring balance equations to be satisfied exactly, Proposition 3.1 suggests that there is a larger class of w such that balance equations are approximately satisfied and $\mathcal{M}^w$ is contractive. Indeed, as we will see below, marginalized operators can recover all contractive multi-step operators as special cases.

3.2 Multi-step off-policy evaluation operators as special cases

We now elucidate the connections between marginalized operators with multi-step off-policy operators. The following result shows that when w is chosen properly, the marginalized operators is equivalent to any given multi-step operator.

Proposition 3.2. Given a multi-step operator $\mathcal{R}^c$ with step-wise trace coefficients c_t, define $w_{x,a}^c(x',a')$ as

$$\frac{1-\gamma}{d_{x,a}^\mu(x',a')}\mathbb{E}_\mu\left[\sum_{t\geq 0}\gamma^t\left(\Pi_{1\leq s\leq t}c_s\right)\mathbb{I}[x_t = x', a_t = a']\right]. \tag{6}$$

If $d_{x,a}^\mu(x',a') = 0$ for some (x',a'), we can instead define $w_{x,a}^c(x',a') = 0$. Let w^c be the matrix form. When $w = w^c$, the two operators are equivalent, $\mathcal{M}^{w^c} = \mathcal{R}^c$.

Proposition 3.2 implies that the space of all contractive marginalized operators contains all contractive multi-step operators. We formally summarize the result as follows.

Corollary 3.3. For any tuple $T = (p,r,\pi,\mu,\gamma)$, Let $\mathcal{C}(T)$ be the space of all step-wise traces (Markovian or non-Markovian) such that $\mathcal{R}^c, c\in\mathcal{C}(T)$ is contractive; let $\mathcal{W}(T)$ be the space of all TD weights such that $\mathcal{M}^w, w\in\mathcal{W}(T)$ is contractive. Then

$$\{\mathcal{R}^c, c\in\mathcal{C}(T)\} \subset \{\mathcal{M}^w, w\in\mathcal{W}(T)\}.$$

As a concrete example of step-wise trace coefficient $c\in\mathcal{C}(T)$, consider the Markovian traces $c_t^{(\text{re})} := \min(\frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}, 1) \leq \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}$ that define the Retrace operators [6]. Let $w^{c^{(\text{re})}}$ be the equivalent marginalized trace. With some algebra, we can show its residual error vector is

$$E_{x,a}^{w^{c^{(\text{re})}}} = \gamma\sum_{t=0}^\infty\gamma^t((P^\pi)^T - \left(P^{\tilde{\pi}}\right)^T)(\left(P^{\tilde{\pi}}\right)^T)^t\delta_{x,a} \geq 0,$$

where $\tilde{\pi}(a|x) := c(x,a)\pi(a|x)$. We can interpret Retrace as imposing an additional yet implicit constraint on c_t, such that the residual error vector is non-negative $E_{x,a}^{w^c} \geq 0$. This is a stronger constraint than requiring the marginalized operator $\mathcal{M}^{w^c}$ to be contractive, which is equivalent to $\eta_{x,a}^{w^c} = (1-\gamma)^{-1}\left\|E_{x,a}^{w^c}\right\|_1 < 1$ as stated in Proposition 5.3. Indeed, as we will see next, by imposing weaker assumptions, marginalized operators contain a larger space of contractive operators than multi-step operators in general.

3.3 Further characterizations of contractive marginalized operators

The above discussion motivates the following question: does the space of contractive marginalized operators contains strictly more elements than contractive multi-step operators? We have the following results.

Proposition 3.4. There exists tuples $T = (p,r,\pi,\mu,\gamma)$ such that either of the following holds

$$\text{(i) } \{\mathcal{R}^c, c\in\mathcal{C}(T)\} \subsetneq \{\mathcal{M}^w, w\in\mathcal{W}(T)\},$$
$$\text{(ii) } \{\mathcal{R}^c, c\in\mathcal{C}(T)\} = \{\mathcal{M}^w, w\in\mathcal{W}(T)\}.$$

Here, we provide some intuitions for case (i). One critical feature of multi-step operators is that the cumulative traces are multiplicative $C_t = (\Pi_{1\leq s\leq t}c_s)$.

Assume a trajectory starting from (x_0, a_0), if the cumulative trace $C_{t^*} = 0$ at some time step t^*, then $C_t = 0, \forall t \geq t^*$. However, by construction, marginalized operators might place TD weights $w_{x_0,a_0}(x_t, a_t)$ such that $w_{x_0,a_0}(x_{t^*}, a_{t^*}) = 0$ and $w_{x_0,a_0}(x_{t'}, a_{t'}) \neq 0$ for some $t' > t^*$. In other words, marginalized operators could *regenerate traces* while multi-step operators cannot. This implies that for such w, there does not exist $c \in \mathcal{C}(T)$ such that $\mathcal{R}^c = \mathcal{M}^w$. We provide specific instances where such phenomenon exist, see Appendix A for the full derivations.

The above result bears important implications to Section 5, where we apply operators $\mathcal{M}^{w_\psi}$ with parameterized TD weights w_ψ. They could be interpreted as directly parameterizing the space of contractive marginalized operators, without necessarily having any multi-step equivalents.

4 Understanding Marginalized Off-Policy Evaluation Operators

We have seen that by properly selecting w, marginalized operators recover multi-step operators as special cases. We provide insights on marginalized operators from a few different perspectives. We start with some background.

4.1 Stochastic estimates of evaluation operators

Since operators are defined in expectations, a naive way to construct stochastic estimates is to directly draw samples from the expectations and compute empirical averages. For example, given a trajectory $(x_t, a_t)_{t=0}^{\infty}$ starting from $x_t = x, a_t = a$, a stochastic estimate to $\mathcal{R}^c Q(x, a)$ is

$$\hat{\mathcal{R}}^c Q(x,a) = Q(x,a) + \sum_{t=0}^{\infty} \gamma^t \left(\Pi_{1 \leq s \leq t} c_s\right) \hat{\Delta}_t,$$

where $\hat{\Delta}_t = r_t + \gamma Q\left(x_{t+1}, \pi(x_{t+1})\right) - Q(x_t, a_t)$ are estimates of Bellman errors. We call this *trajectory based* estimate as the estimate sums over data over the entire trajectory. We could also define a *random time based* estimate with a random time τ such that $P(\tau = n) = (1-\gamma)\gamma^n$ for $n \geq 0$.

$$\hat{\mathcal{R}}_\tau^c Q(x,a) = Q(x,a) + (1-\gamma)^{-1} \left(\Pi_{1 \leq s \leq \tau} c_s\right) \hat{\Delta}_\tau.$$

Both estimates are unbiased. Similarly, we define unbiased stochastic estimates for the marginalized evaluation operators, such that their expectations are $\mathcal{M}^w Q(x, a)$.

$$\hat{\mathcal{M}}^w Q(x,a) = Q(x,a) + \sum_{t=0}^{\infty} \gamma^t w_{x,a}(x_t, a_t) \hat{\Delta}(x_t, a_t),$$

$$\hat{\mathcal{M}}_\tau^w Q(x,a) = Q(x,a) + (1-\gamma)^{-1} w_{x,a}(x_\tau, a_\tau) \hat{\Delta}(x_\tau, a_\tau).$$

4.2 Connections to conditional importance sampling

Interestingly, the conversion of the step-wise trace coefficient c_t into equivalent TD weights $w_(x, a)^c$ as defined in Eqn (6) is closely related to condition importance sampling (IS) [19, 20].

Proposition 4.1. Let τ be an integer-valued random time, such that $P(\tau = n) = (1-\gamma)\gamma^n, \forall n \geq 0$. For any step-wise trace coefficient c_t, its equivalent TD weights $w(x', a')$ is

$$w_{x,a}^c(x', a') = \mathbb{E}_{\mu,\tau}\left[(\Pi_{1 \leq s \leq \tau} c_s) \mid x_\tau = x', a_\tau = a'\right].$$

In other words, $w_{x,a}^c(x', a')$ is the conditional expectation of the random cumulative traces $(\Pi_{1 \leq s \leq \tau} c_s)$ conditional on the event $x_\tau = x', a_\tau = a'$. In general, conditional IS is a useful technique for variance reduction [21], because for any two random variables x, a, $\mathbb{V}[X] \geq \mathbb{V}[\mathbb{E}[X|Y]]$. This implies a variance reduction property of stochastic estimates to the marginalized operators.

Corollary 4.2. Assume that both state transitions and rewards are deterministic. While having the same expectations, the random-time based estimate for the marginalized operator has smaller variance compared to that of the multi-step operator,

$$\mathbb{V}\left[\hat{\mathcal{M}}_\tau^{w^c} Q(x,a)\right] \leq \mathbb{V}\left[\hat{\mathcal{R}}_\tau^c Q(x,a)\right].$$

Importantly, Corollary 4.2 assumes that both state transitions and rewards are deterministic; there is no provable variance reduction when, e.g., the rewards are stochastic. In Appendix B, we graphically present the relations between the four estimates to different operators introduced above.

Remarks on trajectory based estimates. Trajectory based estimates usually have smaller variance than the random time based counterparts. This is because

$$\hat{M}^{w^c} Q(x_0,a_0) = \mathbb{E}\left[\hat{\mathcal{M}}_\tau^{w^c} Q(x_0,a_0) \,\middle|\, (x_t, a_t, r_t)_{t=0}^{\infty}\right],$$

$$\hat{\mathcal{R}}^c Q(x_0,a_0) = \mathbb{E}\left[\hat{\mathcal{R}}_\tau^c Q(x_0,a_0) \,\middle|\, (x_t, a_t, r_t)_{t=0}^{\infty}\right].$$

Though Collorary 4.2 shows the order of variance between random time based estimates, the order of variance of the trajectory based estimates $\hat{\mathcal{R}}^c Q(x, a)$ vs. $\hat{\mathcal{M}}^{w^c} Q(x, a)$ are not clear. Similar results have been observed in [19], where they show that marginalized IS via extended conditional expectations [22] does not necessarily reduce variance. Nevertheless, in practice, estimates to marginalized operators usually reduce variance as evidenced empirically [10].

Trade-off of practical estimates. In practice, TD weights w^c are unknown and need to be estimated $w_\psi \approx w^c$. As a concrete example, consider $c_t = \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}$ and $w^c = w_{x,a}^{\pi,\mu}$. To clarify the trade-off, let $Q \equiv 0$. In this case, $\hat{\mathcal{R}}^c Q(x,a) = \sum_{t\geq 0} \gamma^t (\Pi_{1\leq s\leq t} \frac{\pi(a_s|x_s)}{\mu(a_s|x_s)}) r_t$ [4], which might suffer from high variance due to the product of IS ratios [19]. On the other hand, $\hat{\mathcal{M}}^{w_\psi} Q(x,a) = (1-\gamma)^{-1} \sum_{t=0}^{\infty} w_\psi(x_t,a_t) r_t \approx \hat{\mathcal{M}}^{w^c} Q(x,a)$ where $w_\psi \approx w^c$ is a parametric estimate [10]. As argued in prior work, the latter has lower variance due to marginalized IS but at the cost of the bias in the estimate w_ψ. Overall, moving from the multi-step operator $\hat{\mathcal{R}}^c$ to its estimated marginalized counterpart $\hat{\mathcal{M}}^{w_\psi}$, one trade-offs variance with potential bias due to imperfect estimates of w^c [7]. For general step-wise traces c_t and w^c, this trade-off should still hold. As such, the quality of $w_\psi \approx w^c$ determines the quality of downstream updates. We will discuss in Section 5 how to characterize such effects and estimate w_ψ.

Related work on conditional IS. [20] interprets a large class of off-policy evaluation algorithms as a two-stage process: **(1)** start with an initial estimate; **(2)** compute the conditional IS of the estimate w.r.t. some conditioning variables. State-action pairs (x,a) are popular choices of the conditioning variables, e.g., when applied to marginalized IS [14, 10] and eligibility traces [23]. In this work, we interpret marginalized operators as applying a similar procedure to step-wise traces c_t to derive TD weights w^c.

Extension to V-trace operators. So far, our discussion has focused on off-policy evaluation for Q-functions. By interpreting the TD weights as conditional IS of step-wise traces, we can extend this approach to off-policy evaluation of value functions such as V-trace operators [24]. see Appendix C for detailed results.

4.3 Policy evaluation via linear programs and its connections to contractions

The linear programming (LP) formulation of MDPs [25, 18] is an important framework for policy evaluation, which gives rise to a large number of recent work on marginalized off-policy evaluation (e.g., see [13]). Here, we explore how the notion of contraction is in fact consistent with the LPs. We will see that this offers a new way to interpret LP formulation for policy evaluation, and might pave the way for new algorithms.

Dual LP for policy evaluation. Consider the evaluation of $Q^\pi(x,a)$. Denote $R \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ as the reward vector $R(x,a) = r(x,a)$. We directly start with the dual LP where $d \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ are dual variables. The dual LP for policy evaluation is [18]

$$\begin{cases} \min\ (1-\gamma)^{-1} d^T R \\ (1-\gamma)\delta_{x,a} + \gamma (P^\pi)^T d - d = 0 \end{cases} \tag{7}$$

Since the equality constraints are essentially the balance equations defined in Eqn (4), the single feasible (optimal) solution is $d^* = d_{x,a}^\pi$.

Sequence of relaxed LPs as repeated application of contractive operators. We start by assuming an iterative algorithm, where at iteration t we have access to Q-function estimate $Q_t \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$. At iteration $t+1$, consider the dual LP (Eqn (7)) for each (x,a). We augment its objective function as follows

$$\begin{cases} \min\ Q_t^T \delta_{x,a} + (1-\gamma)^{-1} d^T (R + \gamma (P^\pi)^T Q_t - Q_t) \\ (1-\gamma)\delta_{x,a} + \gamma (P^\pi)^T d - d = 0 \end{cases} \tag{8}$$

Note that the augmented dual LP (Eqn (8)) has the same optimal solution as the original dual LP (Eqn (7)) because both of their feasible region contains only $d_{x,a}^\pi$. Let $\eta \in [0,1)$ be a scalar constant. We relax the constraints of the above dual LP as follows,

$$\begin{cases} \min\ Q_t^T \delta_{x,a} + (1-\gamma)^{-1} d^T (R + \gamma P^\pi Q_t - Q_t) \\ (1-\gamma)\delta_{x,a} + \gamma (P^\pi)^T d - d \leq (1-\gamma) u \\ (1-\gamma)\delta_{x,a} + \gamma (P^\pi)^T d - d \geq -(1-\gamma) u \\ 1^T u \leq \eta, u \geq 0 \end{cases} \tag{9}$$

We name the above relaxed problem $\mathrm{LP}^{(t)}(x,a)$. The feasible region of the relaxed dual LP (Eqn (9)) is expanded into a non-trivial polyhedron $\mathcal{D}_{x,a}$ when $\eta > 0$. Instead of requiring balance equations to hold exactly, violations are allowed and their magnitude is controlled by η. Define $Q_{t+1}(x,a)$ to be the objective value of Eqn (9). The following result relates the sequence of LP objectives to contraction.

Proposition 4.3. The following holds for the sequence of values produced by relaxed LPs,

$$\left\| Q_{t+1} - Q^\pi \right\|_\infty \leq \eta \left\| Q_t - Q^\pi \right\|_\infty .$$

To better understand the above result, note that the feasible region $\mathcal{D}_{x,a}$ effectively characterizes all TD weights w that $\mathcal{M}^w$ is contractive with rate at most η. In particular,

$$\mathcal{D}_{x,a} = \left\{ w_{x,a} \odot d_{x,a}^\mu | \eta_{x,a}^w \leq \eta \right\}$$

where $\odot$ is the element-wise product of vectors. As we show below, the iterative process $Q_t \to Q_{t+1}$ is equivalent to applying contractive operators for policy evaluation

Corollary 4.4. For any (x, a), let $w^*_{x,a} = \frac{d^*}{d^\mu_{x,a}} \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ and d^* is the optimal solution to $\text{LP}^{(t)}(x, a)$, then $\eta^{w^*_{x,a}}_{x,a} \leq \eta$ and

$$Q_{t+1}(x, a) = \mathcal{M}^{w^*_{x,a}} Q_t(x, a).$$

In other words, instead of directly outputting $Q^\pi(x, a)$ by solving $\text{LP}^{(0)}(x, a)$, this iterative algorithm solves relaxed problems and generates a sequence of LP values $Q_t \to Q^\pi$ by implicitly applying operator $\mathcal{M}^{w^*_{x,a}}$.

Related ideas. The idea to reduce solving a single LP into a solving a sequence of relaxed LPs has been explored (e.g., in [26, 27]). They consider the LP for policy optimization, and relax constraints by projecting them onto low-dimensional spaces. This is orthogonal to the box relaxation in Eqn (9).

5 Estimating TD Weights

As previously discussed marginalized operators can achieve variance reduction compared to the equivalent multi-step operators (Corollary 4.2). This poses a practical question: given a multi-step operator $\mathcal{R}c^{\pi,\mu}$, how to find its marginalized equivalent $\mathcal{M}^{w^c}$? In other words, how to estimate w^c from c_t? Given a specific step-wise trace coefficient c_t, we seek an algorithm that estimates the equivalent TD weights $w_\psi \approx w^c_{x,a}, \forall (x, a)$. Throughout the discussion, we focus on Markovian step-wise traces that define Retrace operators $0 \leq c(x_t, a_t) \leq \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}$ [6].

We adapt the TD-learning based method introduced in [10] and derive algorithms to estimating TD weights for generic Markovian step-wise traces. We define $\tilde{\pi}(a|x) := \mu(a|x)c(x, a)$ and a scoring function (also called a critic or discriminator) $\mathbf{q} \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$. Consider the loss function,

$$L(\mathbf{q}, w_\psi) := (1-\gamma)q(x, a) + \mathbb{E}_{(x',a')\sim d^\mu_{x,a}}\left[\Gamma(x', a')\right]. \tag{10}$$

Here, we define $\Gamma(x', a')$ as

$$w(x', a')\left(\gamma\mathbb{E}_{x''\sim p(\cdot|x',a')}\left[q\left(x'', \tilde{\pi}(x'')\right)\right] - q(x', a')\right).$$

We now show a few important properties of the loss function that motivate its use as a practical objective,

Lemma 5.1. Given any two class of scoring functions $\mathcal{Q}_1 \subset \mathcal{Q}_2$, $\max_{\mathbf{q}\in\mathcal{Q}_1} L(\mathbf{q}, w) \leq \max_{\mathbf{q}\in\mathcal{Q}_2} L(\mathbf{q}, w), \forall w$ In addition, the TD weights achieve the global optimal $w^c_{x,a} = \arg\min_w \max_{\mathbf{q}\in\mathcal{Q}} L(\mathbf{q}, w)$ for any $\mathcal{Q}$.

The above result motivates the use of the saddle point optimization objective to search for $w_\psi \approx w^c_{x,a}$. Consider optimizing the following objective jointly with respect to ψ and $\mathbf{q}$,

$$\min_\psi \max_{\mathbf{q}\in\mathcal{Q}} L(\mathbf{q}, w_\psi). \tag{11}$$

The outcome of the optimization ψ can then be used as an approximation $w^\psi \approx w^c_{x,a}$. To characterize the quality of the approximation, note that when $\mathcal{Q}$ contains a large set of scoring functions, the solution ψ^* to Eqn (11) should be closer to $w^c_{x,a}$. This is captured by the following result.

Proposition 5.2. For any sub-probability measure $\tilde{\pi}$, Let $T_{\tilde{\pi}}(x', a'|x, a) := p(x'|x, a)\tilde{\pi}(a'|x')$ be the one-step marginal transition probability. Let $T^t_{\tilde{\pi}}(x', a'|x, a)$ be the t-time composition of $T_{\tilde{\pi}}(\cdot|x, a)$. Given a target state-action pair (x^*, a^*), define the scoring function $q(x, a, x^*, a^*) := \sum_{t\geq 0}\gamma^t T^t_{\tilde{\pi}}(x, a|x^*, a^*)$. Then if $\mathcal{Q}_T(x, a, x^*, a^*) = \{\pm q(x, a, x^*, a^*)\} \subset \mathcal{Q}$, the following holds,

$$|w_\psi(x^*, a^*) - w^c_{x,a}(x^*, a^*)| \leq \frac{\max_{\mathbf{q}\in\mathcal{Q}} L(\mathbf{q}, w_\psi)}{d^\mu_{x,a}(x^*, a^*)}.$$

When $c_t = \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}$, Proposition 5.2 reduces to Theorem 6 in [10] as a special case. In practice, however, it might not be necessary to estimate accurately at each point (x, a). This is because for practical purposes, we only need the downstream operator $\mathcal{M}^{w_\psi}$ to be contractive. The following section discusses how the objective can be directly used for optimizing the contraction rate.

5.1 Optimizing for the contraction rate

So far, we have discussed how to optimize the parameter ψ such that w_ψ matches a particular w^c for some particular trace coefficient c_t. The following result shows that how one could directly minimize the local contraction rate $\eta^{w_\psi}_{x,a}$, without the need to commit to any trace coefficient c_t.

Proposition 5.3. Assume that $\mathcal{Q}_b = \{\pm\delta(x = x^*, a = a^*), \forall (x^*, a^*)\} \subset \mathcal{Q}$. When $c_t = \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}$ and $w^c = w^{\pi,\mu}$, the contraction rate of $\mathcal{M}^{w_\psi}$ is upper bounded as $\eta^{w_\psi}_{x,a} \leq \max_{\mathbf{q}\in\mathcal{Q}} L(\mathbf{q}, w_\psi)$.

Even when the TD weights are not estimated perfectly, the estimated marginalized operator $\mathcal{M}^{w_\psi}$ are still properly defined contractive operators. The above result also implies that in the presence of estimation errors, $\mathcal{M}^{w_\psi}$ could still be contractive even when the TD weights w do not exactly match any weights w^c for any particular trace coefficient c_t. As a result, repeated application of the operator still converges to the correct value. This differs from how prior work interprets imperfect weight estimates (e.g., see [10]) as incurring errors to the final prediction in the offline case.

Line styles	Solid	Dashed	Dashed-dot
# Actions $\lvert\mathcal{A}\rvert$	5	10	20
Horizon T	10	20	30
Off-policy β	0	0.3	0.7
Noise σ	0.1	0.5	1.0
Truncation $\bar{c}$	1	2	5

Table 1: Parameter tables of the chain MDP. This table shows the line styles and their corresponding parameters in Figure 1.

Remarks on $\mathcal{Q}_b$. Compared to $\mathcal{Q}_T(x,a,x^*,a^*)$, $\mathcal{Q}_b$ is much more straightforward to parameterize in practice. For example, consider a neural network f_η which takes (x,a) as input and takes tanh as the output activation: $\tanh(f_\eta(x,a)) \in [-1,1]$. When f_η is expressive enough, it parameterizes the convex hull of $\mathcal{Q}_b$.

Other methods for marginalized estimations. Recently, there is a growing interest in marginalized estimation for off-policy evaluation. Besides TD-learning methods, other notable examples include Fenchel-duality based methods [11, 17, 13] and kernel machines [16]. In Appendix E, we derive a Fenchel-duality based approach to estimating TD weights, which naturally extends the original work [11].

6 Experiments

We start with a few tabular examples to build better understanding of the empirical properties of marginalized operators. For all tabular MDPs, we adopt the tabular representation when learning TD weights. Then we evaluate the potential benefits of marginalized operators when combined with multi-step deep RL algorithms. In this latter case, the TD weights are estimated with function approximations.

6.1 Chain MDP

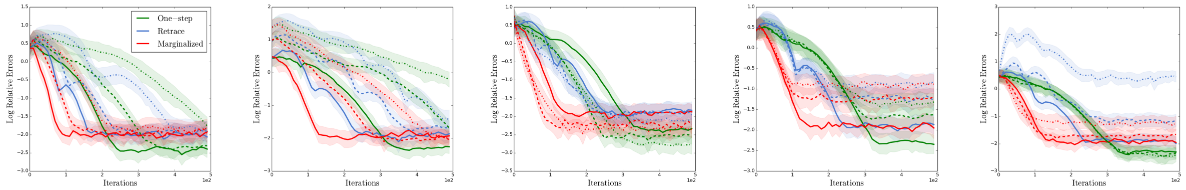
Consider a chain MDP. The reward is zero unless at the rightmost state. At the rightmost state, the reward for action $a \in \mathcal{A}$ is $\mathcal{N}(\mu_a, \sigma^2)$ where $\mu_a = 0$ for all but one action a^* where $\mu_{a^*} = 1$. The episode starts with the leftmost state. For all states, the transition goes to the state to its right with probability 1, no matter what action is taken, until at the rightmost state when the episode terminates. Due to the dynamics of the problem, the episodic horizon is $T \equiv |\mathcal{X}|$. We consider the target policy π as a deterministic policy of choosing action $a = a^*$ at all time. We start with a uniformly random policy u and construct the behavior policy as $\mu = \beta\pi + (1-\beta)u$ where $\beta \in [0,1]$ controls the off-policy level. The problem is on-policy by setting $\beta = 1$. For further details, see Appendix F.

To investigate the impact of different hyper-parameters on the experiment results, we vary the number of actions $|\mathcal{A}|$, the horizon T, the off-policy level β, the noise level σ as they capture different aspects of the MDP. In each sub-plot we vary only one parameter and keep others at the default values. Curves with different line styles correspond to different values of a given parameter, shown in Table 1. The default hyper-parameters of the experiments are in the leftmost column of the table. We compare three baselines: **(1)** one-step operator $\mathcal{T}^\pi Q(x,a) = r(x,a) + \gamma\mathbb{E}[Q(x',\pi(x'))]$, which we recall can be obtained as a special case of Retrace when $\bar{c} = 0$; **(2)** Retrace ($c_t = \min\{\bar{c}, \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}\}$ where $\bar{c} = 1$ by default) and **(3)** marginalized operator $\mathcal{M}^{w_\psi}$ with $w_\psi \approx w^c$ with c_t being the Retrace trace coefficient. Throughout the experiments, we measure the accuracy of the estimate as the relative error $\frac{|\hat{Q}-Q^\pi|}{|Q^\pi|}$, where $\hat{Q}$ is the estimate and Q^π is the ground-truth Q-value.

Results. In Figure 1(a)-(b) shows that the increase in the number of actions or the horizon makes the evaluation more difficult: a large number of actions induces large variance in the estimation due to the increased ratio $\frac{\pi(a|x)}{\mu(a|x)}$; at the same time, long horizons require the propagation of values with more iterations. Overall, the marginalized operator converges faster than Retrace, which further outperforms the one-step operator. In Figure 1(c), we vary the off-policy level: all operators' performance increase as the problem interpolates from very off-policy to near on-policy.

While Figure 1(a)-(c) show the advantages of the marginalized operator, Figure 1(d) highlights potential limitations. As the noise level of the final reward increases, the marginalized operator and Retrace converge to a higher error rate than the one-step operator (similar observations are made in Figure 1(a)-(c)). We speculate that this is because as marginalized estimator and Retrace propagate downstream values more effectively, they also bootstrap noises faster. On the other hand, since the final reward is stochastic, we speculate that one-step operator's incremental backup dampens the variance more significantly, leading to smaller asymptotic errors. The result implies that when there is much noise in the MDP, operators with short bootstrap horizons might be preferred.

To compare Retrace and its marginalized counterpart, we vary the truncation level $\bar{c}$. Here, $\bar{c}$ controls the variance of the target values, as $\bar{c} = 0$ reduces to the one-step operator while $\bar{c} = \infty$ reduces to full importance sampling. As shown in Figure 1(e), the performance of Retrace tends to be unstable when $\bar{c}$ is large; the marginalized operator, converges more stably though the asymptotic errors still increase as $\bar{c}$ increases.



(a) Number of actions (b) Horizon (c) Off-policy level (d) Noise level (e) Truncation level

Figure 1: Comparison of baseline operators on chain MDPs. Each curve is averaged over 100 random seeds. The y-axis shows the evaluation errors, computed as the relative error $\frac{|\hat{Q}-Q^{\pi}|}{|Q^{\pi}|}$ shown in log scale, where $\hat{Q}$ is the estimate and Q^{π} the exact Q-value. The x-axis shows the number of iterations. In each plot, we vary one hyper-parameter of the MDP shown by curves with different line styles. The line styles and their corresponding hyper-parameters are shown in Table 1.

6.2 Open World

We further consider the open world example introduced in [23]. The open world is a deterministic maze with $|\mathcal{X}| = n^2$ states with $n = 10$. At each state, there are four actions $\mathcal{A} = \{\mathrm{L}, \mathrm{U}, \mathrm{R}, \mathrm{D}\}$, each moving the agent to a neighboring state except when moving beyond the boundary, in which case the agent does not move. The agent always starts at the upper left corner. The reward is zero unless the agent transitions into the lower right corner terminal state, where $r = 1$.

We first consider both off-policy evaluation. The agent estimates Q-function tables $\hat{Q}(x, a)$, but in Figure 2 we color-code the value functions for all states computed as $\hat{V}(x) = \sum_a \pi(a|x)\hat{Q}(x, a)$. Here, the behavior policy μ is a uniformly random policy, while the target policy π assigns all probability masses uniformly to $\{\mathrm{D}, \mathrm{R}\}$. We compare three baselines: **(1)** one-step operator; **(2)** Retrace and **(3)** marginalized operator $\mathcal{M}^{w_\psi}$ with $w_\psi \approx w^c$. For further details and more results on policy optimization where off-policy evaluation is used as a subroutine, see Appendix F.

Results. As observed in Figure 2, consistent with results in the chain MDP, the one-step operator propagates information rather slowly compared to the multi-step Retrace. When $\bar{c} = 1$, the performance of Retrace and its marginalized counterpart is highly similar; however, when $\bar{c} = 2$, Retrace becomes unstable. Indeed, moving from lower right to the upper left of the state space, the estimated values do not show any clear trend as in the case of $\bar{c} = 1$, which implies potential divergence. On the other hand, the marginalized operator performs much more stably. All such observations imply that the marginalized operator might achieve an additional effect of variance reduction compared to Retrace. To better interpret the behavior of marginalized operators,we visualize the TD weights w_ψ in Appendix F. The heat maps of the TD weights capture the intuitions of how Bellman errors at future states should impact the estimation at initial states.

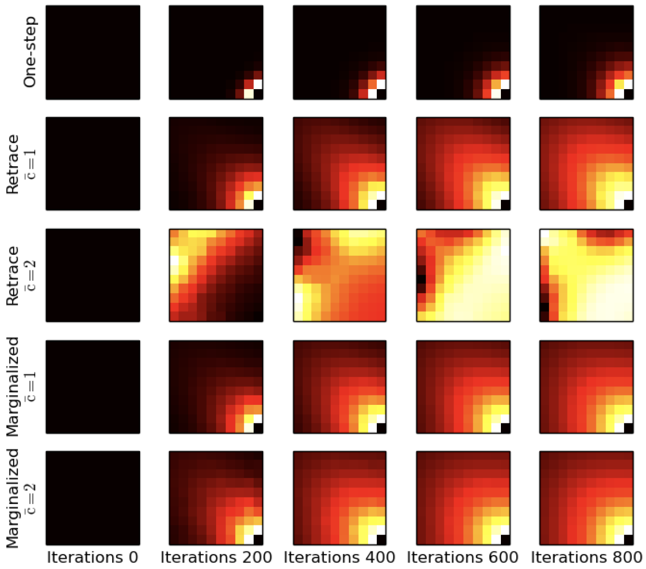
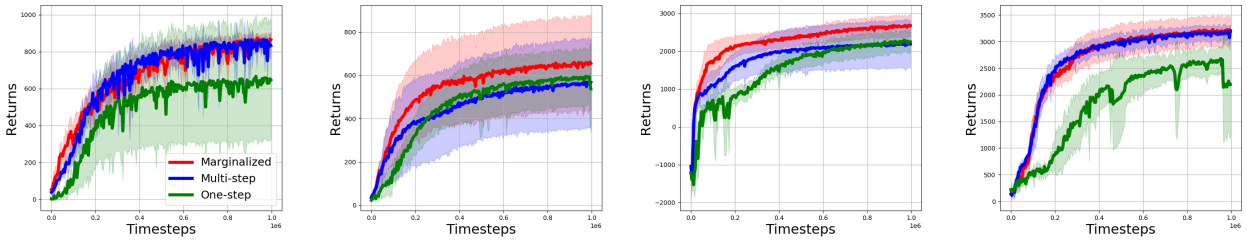


Figure 2: Comparison of operators on the Open World MDP. Each plot is averaged over 100 runs. In each plot, moving from light yellow to red and further to black colors, the estimated values decrease. In Figure 2, going from the leftmost column to rightmost column, the number of iterations increases.

6.3 Deep RL experiments

For high-dimensional state space (or high-dimensional action space), the estimation w_ψ must be combined with more complex function approximation such as neural networks. We use simulated continuous control tasks as the test beds, and compare multi-step RL algorithms against the marginalized counterparts. We consider twin-delayed deep deterministic policy gradient (TD3) [28] as the base algorithm. TD3 implements a deterministic policy $\pi_\phi(x)$ and critic $Q_\theta(x, a)$), both parameterized by neural networks.



(a) **Cheetah(D)** (b) **WalkerRun(D)** (c) **Cheetah(B)** (d) **Ant(B)**

Figure 3: Comparison of operators with deep RL implementations. Each curve is averaged over 5 seeds. The x-axis shows the number of time steps at training time and y-axis shows the evaluation performance. The notation (D) and (B) denote the simulation backends of the testing environments. See Appendix F for further details on the experiments.

The critic is updated by minimizing Bellman errors $\mathbb{E}\left[(Q_\theta(x,a) - Q_{\text{target}}(x,a))^2\right]$ where $Q_{\text{target}}(x,a)$ is constructed by a few alternatives: one-step operator, multi-step operator and its equivalent marginalized operator. Take the one-step operator $\mathcal{T}^\pi$ as an example: given the transition (x,a,r,x') the target is computed as a stochastic estimate to the exact back-up target

$$Q_{\text{target}}(x,a) = r + Q_{\theta^-}(x', \pi(x')),$$

where we can show $\mathbb{E}[Q_{\text{target}}(x,a)] = \mathcal{T}^\pi Q(x,a)$. Here, θ^- is the target network [1]. For multi-step operators, to construct stochastic estimates we follow the procedure in Section 4 to compute $\hat{\mathcal{R}}^c(x,a)$ and $\hat{\mathcal{M}}^w(x,a)$, which are unbiased to $\mathcal{R}^c Q(x,a)$ and $\mathcal{M}^w Q(x,a)$ respectively. In addition to the conventional actor-critic architecture, the marginalized operator also maintains an estimator w_ψ parameterized by a neural network ψ to approximate w^c and applies the resulting operator $\mathcal{M}^{w_\psi}$. See Appendix D for further details on multi-step algorithms and Appendix F for more details on how to estimate w^c with w_ψ in practice with techniques introduced in Section 5.

Results. We show comparison in Figure 3, where we evaluate algorithms over a subset of continuous control tasks [29]. Overall, we find that multi-step updates might outperform or perform similarly as the one-step update, both in terms of learning speed and asymptotic performance; this is consistent with observations made in prior work on multi-step learning for value-based RL algorithms (see, e.g. [30]). Marginalized multi-step updates provide further marginal performance gains over the vanilla multi-step update in terms of the average performance. However, the variations across seeds can be relatively large, indicating that the algorithm might be slightly more unstable due to the online learning of the TD weights.

Limitation of deep RL implementations We discuss some limitations when combining marginalized operators with multi-step deep RL algorithms: training the density ratio estimator w_ψ usually introduces computational overhead and potential instability to the overall algorithm. In large-scale distributed agents (see e.g., [24, 31]), where the data throughput is large, it might not be worthwhile to incur the bias due to the marginalized estimation. It is of interest to further investigate how marginalized estimations can scale to such applications.

7 Conclusion

We have proposed marginalized operators, a general class of off-policy evaluation operators. Marginalized operators bridge the conceptual gap between multi-step operators and marginalized IS methods for off-policy evaluation.

This provides a unified framework to reason about off-policy evaluation, with operator-based approaches and marginalized estimation methods as two seemingly separate yet compatible frameworks, and opens doors to new combinations of algorithmic techniques from both worlds.

One interesting line of future work is to investigate the combination of marginalized operators with traditional multi-step operators, in a similar fashion as how marginalized IS combines with step-wise IS [32].

Acknowledgements. The authors are thankful to constructive comments by anonymous reviewers. The authors thank Zheng Wen for reviewing prior drafts of this paper and are grateful for the supportive colleagues and great research environment at DeepMind.

References

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[2] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[3] Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

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

[5] Anna Harutyunyan, Marc G Bellemare, Tom Stepleton, and Rémi Munos. Q (λ) with off-policy corrections. In *International Conference on Algorithmic Learning Theory (ALT)*, 2016.

[6] Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[7] Mark Rowland, Will Dabney, and Rémi Munos. Adaptive trade-offs in off-policy learning. In *International Conference on Artificial Intelligence and Statistics*, pages 34–44, 2020.

[8] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[9] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.

[10] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366, 2018.

[11] 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*, pages 2318–2328, 2019.

[12] Masatoshi Uehara and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

[13] Ofir Nachum and Bo Dai. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

[14] 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*, pages 9668–9678, 2019.

[15] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*, 2020.

[16] Ali Mousavi, Lihong Li, Qiang Liu, and Denny Zhou. Black-box off-policy estimation for infinite-horizon reinforcement learning. *arXiv preprint arXiv:2003.11126*, 2020.

[17] Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

[18] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

[19] Yao Liu, Pierre-Luc Bacon, and Emma Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. *arXiv preprint arXiv:1910.06508*, 2019.

[20] Mark Rowland, Anna Harutyunyan, Hado Hasselt, Diana Borsa, Tom Schaul, Rémi Munos, and Will Dabney. Conditional importance sampling for off-policy learning. In *International Conference on Artificial Intelligence and Statistics*, pages 45–55. PMLR, 2020.

[21] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[22] Paul Bratley, Bennet L Fox, and Linus E Schrage. *A guide to simulation.* Springer Science & Business Media, 2011.

[23] Hado van Hasselt, Sephora Madjiheurem, Matteo Hessel, David Silver, André Barreto, and Diana Borsa. Expected eligibility traces. *arXiv preprint arXiv:2007.01839*, 2020.

[24] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018.

[25] Daniela Pucci De Farias and Benjamin Van Roy. The linear programming approach to approxi-

mate dynamic programming. *Operations research*, 51(6):850–865, 2003.

[26] Jan Peters, Katharina Mulling, and Yasemin Altun. Relative entropy policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, 2010.

[27] Joan Bas-Serrano, Sebastian Curi, Andreas Krause, and Gergely Neu. Logistic q-learning. *arXiv preprint arXiv:2010.11151*, 2020.

[28] Scott Fujimoto, Herke Van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*, 2018.

[29] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[30] Tadashi Kozuno, Yunhao Tang, Mark Rowland, Rémi Munos, Steven Kapturowski, Will Dabney, Michal Valko, and David Abel. Revisiting peng's q (lambda) for modern reinforcement learning. In *International Conference on Machine Learning*, pages 5794–5804. PMLR, 2021.

[31] Steven Kapturowski, Georg Ostrovski, John Quan, Remi Munos, and Will Dabney. Recurrent experience replay in distributed reinforcement learning. In *International conference on learning representations*, 2018.

[32] Christina Yuan, Yash Chandak, Stephen Giguere, Philip S Thomas, and Scott Niekum. Sope: Spectrum of off-policy estimators. *Advances in Neural Information Processing Systems*, 34, 2021.

[33] Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020.

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

[35] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[36] 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*, pages 5026–5033. IEEE, 2012.

[37] Erwin Coumans. Bullet physics simulation. In *ACM SIGGRAPH 2015 Courses*, page 1. 2015.

[38] Hado V Hasselt. Double q-learning. In *Advances in neural information processing systems*, pages 2613–2621, 2010.

[39] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[40] Joshua Achiam. Spinning Up in Deep Reinforcement Learning. 2018.

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

APPENDICES: Marginalized Operators for Off-Policy Reinforcement Learning

A Proofs

Proposition 3.1. For any TD weights w, the Q-function Q^π is a solution to the fixed point equation $\mathcal{M}^w Q = Q$. For any $Q_1, Q_2 \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$,

$$|\mathcal{M}^w Q_1(x,a) - \mathcal{M}^w Q_2(x,a)| \leq \eta^w_{x,a} \left\|Q_1 - Q_2\right\|_\infty .$$

Let $\delta_{x,a} \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ be the one-hot encoding of (x,a) and let $d^w_{x,a} \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ such that $d^w_{x,a}(x',y') = w_{x,a}(x',y')d^\mu_{x,a}(x',a')$. Then define the *residual error vector*

$$E^w_{x,a} = (1-\gamma)\delta_{x,a} + \gamma(P^\pi)^T d^w_{x,a} - d^w_{x,a},$$

which characterizes how $d^w_{x,a}$ satisfies the *balance equations*

$$(1-\gamma)\delta_{x,a} + \gamma(P^\pi)^T d - d = 0. \tag{4}$$

The local contraction rate is expressed as

$$\eta^w_{x,a} = (1-\gamma)^{-1}\left\|E^w_{x,a}\right\|_1 . \tag{5}$$

The above implies that the operator is contractive when $\max_{x,a}\left\|E^w_{x,a}\right\|_1 < 1-\gamma$.

Proof. We adopt the matrix notation to prove the result. See the proof for Proposition 3.2 for a detailed discussion on the matrix notation as well. By the definition of marginalized operators, define $W \in \mathbb{R}^{(\mathcal{X}\times\mathcal{A})\times(\mathcal{X}\times\mathcal{A})}$ as the tensor such that $W(x,a,x',a') = w^c_{x,a}(x',a')$. Now, we can rewrite for any $Q \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$,

$$\mathcal{M}^w Q = Q + \left[(I-\gamma P^\mu)^{-1} \odot W\right](R + \gamma P^\pi Q - Q),$$

where $\odot$ denotes the element-wise product of two tensors with the same shape. Let $D \coloneqq (I-\gamma P^\mu)^{-1} \odot W$. For any Q_1, Q_2,

$$\mathcal{M}^w Q_1 - \mathcal{M}^w Q_2 = (\gamma D P^\pi - D + I)(Q_1 - Q_2).$$

Now, examine the (x,a)-th component of the vector $\mathcal{M}^w Q_1 - \mathcal{M}^w Q_2$. Through inspection, we can identify $d^w_{x,a}$ as the (x,a)-th row of D scaled by $(1-\gamma)$. This implies

$$\mathcal{M}^w Q_1(x,a) - \mathcal{M}^w Q_2(x,a) = (1-\gamma)^{-1}\left((1-\gamma)\delta_{x,a} + \gamma(P^\pi)^T d^w_{x,a} - d^w_{x,a}\right)^T (Q_1 - Q_2).$$

If we define $E^w_{x,a} \coloneqq (1-\gamma)\delta_{x,a} + \gamma(P^\pi)^T d^w_{x,a} - d^w_{x,a}$, the above rewrites as

$$\mathcal{M}^w Q_1(x,a) - \mathcal{M}^w Q_2(x,a) = (1-\gamma)^{-1}E^w_{x,a}(Q_1 - Q_2).$$

Hence the contraction rate of the operator is $\max_{x,a}(1-\gamma)^{-1}\left\|E^w_{x,a}\right\|_1$.

□

Proposition 3.2. Given a multi-step operator $\mathcal{R}^c$ with step-wise trace coefficients c_t, define $w^c_{x,a}(x',a')$ as

$$\frac{1-\gamma}{d^\mu_{x,a}(x',a')}\mathbb{E}_\mu\left[\sum_{t\geq 0}\gamma^t\left(\Pi_{1\leq s\leq t}c_s\right)\mathbb{I}[x_t = x', a_t = a']\right]. \tag{6}$$

If $d^\mu_{x,a}(x',a') = 0$ for some (x',a'), we can instead define $w^c_{x,a}(x',a') = 0$. Let w^c be the matrix form. When $w = w^c$, the two operators are equivalent, $\mathcal{M}^{w^c} = \mathcal{R}^c$.

Proof. We start by assuming $d_{x,a}^{\mu}(x',a') > 0$ for all $(x,a),(x',a')$. We introduce matrix notations for the marginalized operator. For TD weights w, let W be a matrix such that $W(x,a,x',a') = w_{x,a}(x',a')$. For any two matrices A, B of the same shape, let $A \odot B$ be the element-wise product. Let $R \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ be the expected reward vector such that $R(x,a) = \bar{r}(x,a)$. By the definition of marginalized operators, we rewrite

$$\mathcal{M}^w Q = Q + \left[(I - \gamma P^{\mu})^{-1} \odot W\right](R + \gamma P^{\pi} Q - Q).$$

We first assume that the multi-step operator adopts Markovian step-wise traces. Let $P^{c\mu}$ be the transition matrix defined by the sub-probability measure $c\mu$ such that $P^{c\mu}(x,a,x',a') = p(x'|x,a)\mu(a'|x')c(x',a')$. We can write [6]

$$\mathcal{R}^c Q = Q + (I - \gamma P^{c\mu})^{-1}(R + \gamma P^{\pi} Q - Q).$$

By letting $\mathcal{M}^w = \mathcal{R}^c$, we can see the following is a solution to w

$$W = (I - \gamma P^{c\mu})^{-1} / (I - \gamma P^{\mu})^{-1}. \tag{12}$$

Here, for two matrices A, B of the same shape, we define A/B to be the element-wise division, where it is required that all entries of B are strictly positive. Note that $(I - \gamma P^{c\mu})^{-1} = \sum_{t=0}^{\infty} (\gamma P^{c\mu})^t$ and $(I - \gamma P^{\mu})^{-1} = \sum_{t=0}^{\infty} (\gamma P^{\mu})^t$. This implies that the (x,a,x',a') component of $(I - \gamma P^{\mu})^{-1}$ is $(1-\gamma)^{-1} d_{x,a}^{\mu}(x',a')$, and the (x,a,x',a') component of $(I - \gamma P^{c\mu})^{-1}$ is accordingly

$$\frac{1-\gamma}{d_{x,a}^{\mu}(x',a')} \mathbb{E}_{\mu}\left[\sum_{t\geq 0} \gamma^t \left(\Pi_{1\leq s\leq t} c_s\right) \mathbb{I}[x_t = x', a_t = a'] \,\middle|\, x_0 = x, a_0 = a\right].$$

By reading off components from the matrix equality Eqn (12), we arrive at the desired result.

When the traces are non-Markovian, the proof can be extended naturally. Let $c_t \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ be a vector such that $c_t(x,a)$ defines the step-wise trace at time t after starting with (x,a). The multi-step operator can be written as

$$\mathcal{R}^c Q = Q + \sum_{t=0}^{\infty} \left(\Pi_{0\leq s\leq t} P^{c_t \mu}\right)(R + \gamma P^{\pi} Q - Q). \tag{13}$$

We then arrive at the following sufficient condition for $\mathcal{M}^w = \mathcal{R}^c$

$$W = \sum_{t=0}^{\infty} \left(\Pi_{0\leq s\leq t} P^{c_t \mu}\right) / (I - \gamma P^{\mu})^{-1}.$$

By reading off components of both sides, we arrive at the desired conclusion.

Now in case for some (x,a,x',a'), $d_{x,a}^{\mu}(x',a') = 0$, we can safely set $w_{x,a}^{c}(x',a') = 0$. This is because $d_{x,a}^{\mu}(x',a') = 0$ implies that there is zero probability that the agent arrives in (x',a') starting from (x,a), which means Bellman errors starting from (x',a') are never computed as part of expectation which defines the operator.

Technical conditions for the summation in Eqn (6). It is clear that there exists some step-wise traces c_t such that the summation in Eqn (6) does not converge, e.g., $c_t = \frac{1}{\gamma}$. We impose a condition: (**C.1**) The step-wise traces c_t should be such that $\mathcal{R}^c Q$ is finite under the definition in Eqn 13. Naturally, (**C.1**) implies that $\sum_{t=0}^{\infty} \left(\Pi_{0\leq s\leq t} P^{c_t \mu}\right)$ is finite element-wise, which further implies that the infinite sum $\mathbb{E}_{\mu}\left[\sum_{t\geq 0} \gamma^t \left(\Pi_{1\leq s\leq t} c_s\right) \mathbb{I}[x_t = x', a_t = a'] \,\middle|\, x_0 = x, a_0 = a\right]$ is finite for all $(x,a),(x',a')$. Note that (**C.1**) is very weak and is valid for all situations of interest to us.

□

Corollary 3.3. For any tuple $T = (p, r, \pi, \mu, \gamma)$, Let $\mathcal{C}(T)$ be the space of all step-wise traces (Markovian or non-Markovian) such that $\mathcal{R}^c, c \in \mathcal{C}(T)$ is contractive; let $\mathcal{W}(T)$ be the space of all TD weights such that $\mathcal{M}^w, w \in \mathcal{W}(T)$ is contractive. Then

$$\{\mathcal{R}^c, c \in \mathcal{C}(T)\} \subset \{\mathcal{M}^w, w \in \mathcal{W}(T)\}.$$

Proof. Given any step-wise traces c_t (Markovian or non-Markovian), we can compute corresponding marginalzied traces w via Eq (6). Then $\mathcal{R}^c = \mathcal{M}^w$ per Proposition 3.2. This implies the desired result in the corollary. $\square$

Proposition 3.4. There exists tuples $T = (p, r, \pi, \mu, \gamma)$ such that either of the following holds

$$\text{(i) } \{\mathcal{R}^c, c \in \mathcal{C}(T)\} \subsetneq \{\mathcal{M}^w, w \in \mathcal{W}(T)\},$$
$$\text{(ii) } \{\mathcal{R}^c, c \in \mathcal{C}(T)\} = \{\mathcal{M}^w, w \in \mathcal{W}(T)\}.$$

Proof. We start with some clarifications on notations. The TD weights c_t could be either Markovian or non-Markovian. In the latter case, we require that c_t is measurable w.r.t. $(x_s, a_s)_{s \leq t}$. Given a tuple of MDP, policy and discount factor $T = (r, p, \pi, \mu, \gamma)$, Note that here c_t could be Markovian or non-Markovian. Let $\mathcal{C}_{\text{markov}}(T) \in \mathbb{R}^{(\mathcal{X}\times\mathcal{A})\times(\mathcal{X}\times\mathcal{A})}$ be the set of Markovian traces such that $\mathcal{R}^c$ is contractive; let $\mathcal{C}_{\text{non-markov}}(T) \in \left(\mathbb{R}^{\mathcal{X}\times\mathcal{A}}\right)^H$ be the set of non-Markovian traces such that $\mathcal{R}^c$ is contractive, where H is horizon of the Markov chain induced by π starting from any state-action pair. In general, we consider $H = \infty$. As such, for any T, $\mathcal{C}(T) = \mathcal{C}_{\text{markov}}(T) \cup \mathcal{C}_{\text{non-markov}}(T)$. Finally, let $\mathcal{W}(T)$ be the set of TD weights such that for any $w \in \mathcal{W}(T)$, any $\mathcal{M}^w \in \mathcal{W}(T)$ is contractive.

Per Proposition 3.2, we can start with any $c \in \mathcal{C}(T)$ and project it into a $w \in \mathcal{W}(T)$. For convenience of the discussion, we denote such a projection as $f_{c\to w}^T$, where the T denotes that this projection generally depends on T (e.g., the expectation defined in Eqn (6) is computed with respect to the dynamics p). Formally, we can write $f_{c\to w}^T : \mathcal{C}(T) \mapsto \mathcal{W}(T)$.

We state a few important properties of $f_{c\to w}^T$ as lemmas.

Lemma A.1. When constrained $f_{c\to w}^T$ to Markovian traces, let the constrained mapping be $f_{c\to w}^{\mathcal{C}_{\text{markov}},T} : \mathcal{C}_{\text{markov}}(T) \mapsto W(T)$. There exists tuples T such that $f_{c\to w}^{\mathcal{C}_{\text{markov}},T}$ is not surjective.

Proof. We prove by constructing a counterexample where for some T, there exists a $w \in \mathcal{W}(T)$ that cannot be obtained by first picking a Markovian trace $c \in \mathcal{C}_{\text{markov}}$ and then project it through $f_{c\to c}^{\mathcal{C}_{\text{markov}},T}$.

Consider a deterministic chain MDP with N states $\{x_i\}_{i=1}^N$. All first $N-1$ states transition deterministically to the next state on the right. The last (rightmost) state is absorbing. Assume also $\pi = \mu$ to be both deterministic policy. Consider the TD weights w^* such that its (x, a, x', a') component is $w_{x,a}(x', a') = \frac{\delta_{x'=x,a'=a}}{d_{x,a}^{\mu}(x',a')}$. In this case, the operator $\mathcal{M}^{w^*}$ is exactly the one-step TD operator. Starting from state $x_i, 1 \leq i \leq N-1$, the marginalized operator is

$$\mathcal{M}^w Q(x_i, a_i) = Q(x_i, a) + \left(r_i + \gamma Q\left(x_{i+1}, \pi(x_{i+1})\right) - Q(x_i, a)\right).$$

The step-wise operator is

$$\mathcal{R}^c Q(x_i, a_i) = Q(x_i, a_i) + \sum_{i \leq j \leq N-1} \gamma^{j-i} \left(c_{i+1} ... c_j\right) \left(r_j + \gamma Q\left(x_{j+1}, \pi(x_{j+1})\right) - Q(x_j, a_j) - Q(x_j)\right) + F(N),$$

where $F(N)$ is some function of the last state. Now, we find c such that $\mathcal{M}^{w^*} = \mathcal{R}^c$. By matching coefficients of the term $Q(x_{i+1}, a_{i+1})$, it is necessary that $c(x_i, a_i) = 1$. However, by setting $c(x_i, a_i) = 1$, $\mathcal{R}^c \neq \mathcal{M}^{w^*}$. In other words, there does not exist a Markovian trace $c \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ such that $f_{c\to w}(c) = w^*$. This implies that under this setup, the mapping is not surjective. $\square$

Lemma A.2. Let $W^+(T) = \{w \in W(T), w > 0\} \subset W(T)$. For any T, $f_{c\to w}^T$ is surjective to $W^+(T)$.

Proof. Intuitively, for those TD weights w that could not be realized by Markovian step-wise traces, we need to construct non-Markovian step-wise traces c_t to construct them, such that $f_{c\to w}^T(c) = w$.

We construct non-Markovian step-wise traces as follows. Given $w \in \mathcal{W}^+(T)$. starting from (x, a), the step-wise coefficient at time $t \geq 0$ is computed as

$$c_t \coloneqq \frac{w_{x,a}(x_t, a_t)}{w_{x,a}(x_{t-1}, a_{t-1})},$$

where we define $w_{x,a}(x_t, a_t) = 1$ for $t = -1$. We can show that by such a construction, $(\Pi_{1 \leq s \leq t} c_s) = w_{x,a}(x_t, a_t)$ and as such

$$f_{c \to w}(T)(c) = w.$$

□

Lemma A.3. There exists T, such that $f_{c \to w}(T)$ is **not** surjective to $W(T)$.

Proof. We construct a counterexample of T. In this case, we seek TD weights $w \in W(r, p, \pi, \mu, \gamma)$ such that we cannot find $c \in \mathcal{C}(T)$ such that $f_{c \to w}(T)(c) = w$. Notably, in this case, $\mathcal{C}(T)$ should contain all step-wise traces, both Markovian and non-Markovian ones.

Consider a deterministic chain MDP with $|\mathcal{X}| = N = 5$ states $\{x_i\}_{i=1}^N$ and $|\mathcal{A}| = 2$ actions $\{a_i\}_{i=1}^2$. All first $N - 1$ states transition deterministically to the next state on the right. The last (rightmost) state is absorbing. Assume that $\pi = \mu$ are both uniformly random. Finally, let $\gamma = 0.8$.

Consider the contraction property of $\mathcal{M}^w$ starting from the state (x_1, a_1). We can show that by defining $d_{x_1,a_1}(x', a') = 0$ except

$$d_{x_1,a_1}(x_1, a_1) = 0.2, d_{x_1,a_1}(x_3, a_3) = 0.01.$$

Then we set $w_{x_1,a_1} = \frac{d_{x_1,a_1}}{d^{\mu}_{x_1,a_1}}$ (element-wise division). We can show that

$$|\mathcal{M}^w Q_1 - \mathcal{M}^w Q_2|(x_1, a_1) \leq 0.89 \left\|Q_1 - Q_2\right\|_\infty.$$

This implies that the resulting operator $\mathcal{M}^w$ is contractive for the pair (x_1, a_1). We can complete the definition of w for other state-action pairs (x, a) by specifying $w_{x,a}$ properly. Concretely, as an example, we might set $w_{x,a}^{x',a'} = \frac{\delta_{x'=x,a'=a}}{d^{\mu}_{x,a}(x',a')}$ so that $|\mathcal{M}^w Q_1 - \mathcal{M}^w Q_2|(x', a') \leq \gamma \left\|Q_1 - Q_2\right\|_\infty = 0.8 \left\|Q_1 - Q_2\right\|_\infty$ for any $(x', a') \neq (x_1, a_1)$. Overall, the operator is contractive

$$\left\|\mathcal{M}^w Q_1 - \mathcal{M}^w Q_2\right\|_\infty \leq 0.89 \left\|Q_1 - Q_2\right\|_\infty.$$

Now, we argue why this particular choice of w_{x_1,a_1} cannot be realized by any step-wise traces. Note that since by construction, $d_{x_2,a} = 0, \forall a \in \{a_1, a_2\}$. This implies that starting from (x_1, a_1), if we seek any step-wise traces which are equivalent to d_{x_1,a_1}, they must cut the traces at (x_2, a). A direct consequence of this result is that $c(x_2, a) = 0$ for both Markovian or non-Markovian traces. However, since the traces are multiplicative, this further means that the cumulative product of traces at (x_3, a) would be zero. This does not replicate the behavior of d_{x_1,a_1}, whose entry at (x_3, a_3) is constructed to be $0.01 > 0$.

To summarize, the above example shows that under this particular set of T, there exists a w that cannot be realized by any step-wise traces through the mapping $f_{c \to w}(T)$. Hence the result is concluded. □

Lemma A.4. There exists T, such that $f_{c \to w}(T)$ is surjective to $W(T)$.

Proof. Consider a special case where we have $|\mathcal{X}| = 2$ states and $|\mathcal{A}| = 1$ action. Let x_1, x_2 be the states and a_1 the single action. Assume also all rewards are deterministic. As such, the policy π, μ are trivial as $\pi(a_1|x) = \mu(a_1|x) = 1, \forall x$. The transition matrix is

$$P^\pi = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

With the above setup, consider any marginalized trace at (x_1, a_1), $w_{x_1,a_1} \in W(r, p, \pi, \mu, \gamma)$. Note that $w_{x_1,a_1} \in \mathbb{R}^2$. Let $c_t, t \geq 1$ be the non-Markovian step-wise trace starting from (x_1, a_1). Define the one-step Bellman errors $\Delta_1 \coloneqq r_1 + \gamma Q(x_2, a_1) - Q(x_1, a_1), \Delta_2 \coloneqq r_2 + \gamma Q(x_1, a_1) - Q(x_2, a_1)$.

The marginalized operator evaluated at (x_1, a_1) is

$$\mathcal{M}^w Q(x_1, a_1) = Q(x_1, a_1) + \left(1 + \gamma^2 + \gamma^4 + ...\right) w_{x_1,a_1}^{x_1,a_1} \Delta_1 + \left(\gamma + \gamma^3 + \gamma^5 + ...\right) w_{x_1,a_1}^{x_2,a_1} \Delta_2.$$

The step-wise operator is

$$\mathcal{R}^c Q(x_1,a_1) = Q(x_1,a_1) + \left(1+\gamma^2 c_1 c_2 + \gamma^4 c_1 c_2 c_3 c_4 + ...\right)\Delta_1 + \left(\gamma c_1 + \gamma^3 c_1 c_2 c_3 + ...\right)\Delta_2.$$

We can identify the following solution c to satisfy the equality $\mathcal{M}^w Q(x_1,a_1) = \mathcal{R}^c(x_1,a_1)$.

$$c_1 = 1, c_2 = \frac{A-1}{\gamma^2}, c_3 = B - \gamma - \gamma(A-1), c_4 = c_5 = ... = 0,$$

where $A = \left(1+\gamma^2+\gamma^4\right) w_{x_1,a_1}(x_1,a_1), B = \left(\gamma+\gamma^3+\gamma^5+...\right) w_{x_1,a_1}(x_2,a_1)$. Note that the solution always exists regardless of w_{x_1,a_1}. In a similar way, we can solve for non-Markovian traces for w_{x_2,a_1} as well. We conclude for any w, there exists non-Markovian traces c such that $f_{c\to w}(r,p,\pi,\mu,\gamma)(c) = w$ for the above (r,p,π,μ,γ). □

The above lemmas characterize the space of $\{\mathcal{R}^c, c \in \mathcal{C}(T)\}$ relative to $\{\mathcal{M}^w, w \in \mathcal{W}(T)\}$. From Lemma A.3 we conclude case (i) of the proposition; from Lemma A.4, we conclude the case (ii) of the proposition.

□

Proposition 4.1. Let τ be an integer-valued random time, such that $P(\tau = n) = (1-\gamma)\gamma^n, \forall n \geq 0$. For any step-wise trace coefficient c_t, its equivalent TD weights $w(x',a')$ is

$$w^c_{x,a}(x',a') = \mathbb{E}_{\mu,\tau}\left[(\Pi_{1\leq s\leq\tau} c_s) \mid x_\tau = x', a_\tau = a'\right].$$

Proof. The definition of w^c could rewrite as

$$w^c_{x,a}(x',a') \cdot d^\mu_{x,a}(x',y') = \mathbb{E}_{\mu,\tau}\left[(\Pi_{1\leq s\leq\tau}\mathbb{I}[x_\tau = x', a_\tau = a']\right]$$

As such, we expand the RHS of the above

$$\begin{aligned} w^c_{x,a}(x',a') \cdot d^\mu_{x,a}(x',y') &= \mathbb{E}_{\mu,\tau}\left[(\Pi_{1\leq s\leq\tau}\mathbb{I}[x_\tau = x', a_\tau = a']\right] \\ &= \mathbb{E}_{\mu,\tau}\left[(\Pi_{1\leq s\leq\tau} c_s) \mid x_\tau = x', a_\tau = a'\right] \\ &\quad \times P_\mu(x'_\tau = x', a_\tau = a | x_0 = x, a_0 = a). \end{aligned} \tag{14}$$

Also note that $d^\mu_{x,a}(x',a') \coloneqq (1-\gamma)\sum_{t\geq 0}\gamma^t P_\mu(x_t = x', a_t = a' | x_0 = x, a_0 = a) = P_\mu(x_\tau = x', a_\tau = a' | x_0 = x, a_0 = a)$, which cancel on both sides of the equation. Hence we conclude the equality. □

Corollary 4.2. Assume that both state transitions and rewards are deterministic. While having the same expectations, the random-time based estimate for the marginalized operator has smaller variance compared to that of the multi-step operator,

$$\mathbb{V}\left[\hat{\mathcal{M}}^{w^c}_\tau Q(x,a)\right] \leq \mathbb{V}\left[\hat{\mathcal{R}}^c_\tau Q(x,a)\right].$$

Proof. With Proposition 4.1, we have $w^c(x_\tau,a_\tau) = \mathbb{E}_{\mu,\tau}\left[(\Pi_{1\leq s\leq\tau} c_s) \mid x_\tau, a_\tau\right]$. Further,

$$\begin{aligned} w(x_\tau,a_\tau)\hat{\Delta}^\pi_\tau &= \mathbb{E}_{\mu,\tau}\left[(\Pi_{1\leq s\leq\tau} c_s) \mid x_\tau, a_\tau\right]\hat{\Delta}^\pi_\tau \\ &= \mathbb{E}_{\mu,\tau}\left[(\Pi_{1\leq s\leq\tau} c_s)\delta^\pi_\tau \mid x_\tau, a_\tau\right] \end{aligned}$$

Note that since the transitions are deterministic $\hat{\Delta}^\pi_\tau$ is a measurable function of (x_τ, a_τ) and could be taken out of the expectation. Then with the tower property of variance $\mathbb{V}[X] \geq \mathbb{V}[\mathbb{E}[X \mid Y]]$, by letting $X = (\Pi_{1\leq s\leq\tau} c_s)\hat{\Delta}^\pi_\tau$ and $Y = (x_\tau, a_\tau)$ we conclude the result. □

Proposition A.5. For any step-wise trace coefficient c_t, its equivalent TD weights w^c and $d^{w^c}_{x,a} \coloneqq d^\mu_{x,a} \odot w^c_{x,a}$,

$$d^{w^c}_{x,a} = (1-\gamma)\delta_{x,a} + \gamma(P^{\tilde{\pi}})^T d^{w^c}_{x,a}, \tag{15}$$

where $\tilde{\pi}(a|x) = \pi(a|x)c(x,a)$ and $\tilde{\pi}(a|x) \coloneqq \mu(a|x)c(x,a)$ is a non-negative measure for any $0 \leq c(x,a) \leq \frac{\pi(a|x)}{\mu(a|x)}$.

Proof. We show the Bellman equation directly from the definition of $d_{x,a}^{w^c}(x',a')$. In the following, we always condition on $x_0 = x, a_0 = a$ inside expectations. For the simplicity of notations, we drop this conditioner by default. It is clear that by construction,

$$d_{x,a}^{w^c}(x',a') = (1-\gamma)\mathbb{E}_\mu\left[\sum_{t\geq 0}\gamma^t(\Pi_{s=1}^t c_s)\mathbb{I}[x_t = x', a_t = a']\right]$$

We rewrite the above into the following

$$\begin{aligned} d_{x,a}^{w^c}(x',a') &= (1-\gamma)\mathbb{I}[x_0 = x', a_0 = a'] + \mathbb{E}_\mu\left[\sum_{t\geq 1}\gamma^t\left(\Pi_{s=1}^t c_s\right)\mathbb{I}[x_t = x', a_t = a']\right] \\ &= (1-\gamma)\mathbb{I}[x_0 = x', a_0 = a'] + \gamma\mathbb{E}_\mu\left[\sum_{u\geq 0}\gamma^u\left(\Pi_{s=1}^u c_s\right)c_{u+1}\mathbb{I}[x_{u+1} = x', a_{u+1} = a']\right], \end{aligned}$$

where in the second equality we apply the transformation $u = t-1$. Now, let $h_u \coloneqq \{x_0 = x, a_0, ...x_u, a_u\}$ denote the sequence of random variables until time u. For each term in the summation, for any given $u \geq 0$,

$$\begin{aligned} \mathbb{E}_\mu\left[\gamma^u(\Pi_{s=1}^u c_s)c_{u+1}\mathbb{I}[x_{u+1} = x', a_{u+1} = a']\right] &= \sum_{y\in\mathcal{X}, b\in\mathcal{A}}\mathbb{E}_\mu\left[\gamma^u(\Pi_{s=1}^u c_s)c_{u+1}\mathbb{I}[x_{u+1} = x', a_{u+1} = a']\mathbb{I}[x_u = y, a_u = b]\right] \\ &= \sum_{y\in\mathcal{X}, b\in\mathcal{A}}\mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\gamma^u(\Pi_{s=1}^u c_s)c_{u+1}\mathbb{I}[x_{u+1} = x', a_{u+1} = a']\mathbb{I}[x_u = y, a_u = b] \mid h_u\right]\right] \\ &= \sum_{y\in\mathcal{X}, b\in\mathcal{A}}\mathbb{E}_\mu\left[\gamma^u(\Pi_{s=1}^u c_s)\mathbb{I}[x_u = y, a_u = b]P^{\tilde{\pi}}(x_u, a_u, x', a')\right] \\ &= \mathbb{E}_\mu\left[\gamma^u(\Pi_{s=1}^u c_s)\mathbb{I}[x_u = y, a_u = b]P^{\tilde{\pi}}(y, b, x', a')\right]. \end{aligned}$$

In the above, we have used the equality,

$$\mathbb{E}_\mu\left[c_{u+1}\mathbb{I}[x_{u+1} = x', a_{u+1} = a'] \mid h_u\right] = \mathbb{E}_\mu\left[c_{u+1}\mathbb{I}[x_{u+1} = x', a_{u+1} = a'] \mid x_u, a_u\right] = P^{\tilde{\pi}}(x_u, a_u, x', a'),$$

which derives from the definition of the transition matrix. Finally, we sum up over the time step k to yield the final fixed point equation,

$$d_{x,a}^{w^c}(x',a') = (1-\gamma)\mathbb{I}[x_0 = x', a_0 = a'] + \gamma\sum_{y\in\mathcal{X}, b\in\mathcal{A}}d_{x,a}^{w^c}(y,b)P^{\tilde{\pi}}(y,b,x',a').$$

By rewriting the above equation into the matrix form, we conclude the proof.

Alternative proof by matrix notations. We can derive much simpler alternative proof with matrix notations. Let $d^{w^c} \in \mathbb{R}^{(\mathcal{X}\times\mathcal{A})\times(\mathcal{X}\times\mathcal{A})}$ be a matrix such that $d^{w^c}(x,a,x',a') = d_{x,a}^{w^c}(x',a')$. Also define the visitation distribution matrix $d^\mu = (1-\gamma)(I-\gamma P^\mu)^{-1}$. Recall that from the proof of Proposition 3.2, in matrix form,

$$W = (I-\gamma P^{c\mu})^{-1}/(I-\gamma P^\mu)^{-1}.$$

Then by construction,

$$d^{w^c} = (1-\gamma)W\odot d^\mu = (1-\gamma)(I-\gamma P^{c\mu})^{-1} = (1-\gamma)\sum_{t=0}^{\infty}(\gamma P^{c\mu})^t.$$

Then naturally, d^{w^c} satisfies the following Bellman equations,

$$d^{w^c} = (1-\gamma) + \gamma P^{c\mu}d^{w^c}.$$

When indexing the row at (x,a), we arrive at the desired result.

□

Proposition 5.2. For any sub-probability measure $\tilde{\pi}$, Let $T_{\tilde{\pi}}(x',a'|x,a) \coloneqq p(x'|x,a)\tilde{\pi}(a'|x')$ be the one-step marginal transition probability. Let $T^t_{\tilde{\pi}}(x',a'|x,a)$ be the t-time composition of $T_{\tilde{\pi}}(\cdot|x,a)$. Given a target state-action pair (x^*,a^*), define the scoring function $q(x,a,x^*,a^*) \coloneqq \sum_{t\geq 0}\gamma^t T^t_{\tilde{\pi}}(x,a|x^*,a^*)$. Then if $\mathcal{Q}_T(x,a,x^*,a^*) = \{\pm q(x,a,x^*,a^*)\} \subset \mathcal{Q}$, the following holds,

$$|w_\psi(x^*,a^*) - w^c_{x,a}(x^*,a^*)| \leq \frac{\max_{\mathbf{q}\in\mathcal{Q}} L(\mathbf{q},w_\psi)}{d^\mu_{x,a}(x^*,a^*)}.$$

Proof. Define $d_\psi \coloneqq w_\psi \odot d^\mu_{x,a}$. By construction of the objective $L(\mathbf{q},w_\psi)$, we can rewrite the objective as an inner product,

$$L(\mathbf{q},w_\psi) = \mathbf{q}^T[(1-\gamma)\delta_{x,a} + \gamma(P^{\tilde{\pi}})^T d_\psi - d_\psi], \tag{16}$$

where $\tilde{\pi}(a|x) = \mu(a|x)c(x,a)$ is a sub-probability measure. Per results in Proposition A.5, the objective satisfies the following equation when the second argument is w^c

$$L(\mathbf{q},w^c) = 0.$$

Hence, we can rewrite Eqn (16) as the following

$$\begin{aligned} L(\mathbf{q},w_\psi) &= L(\mathbf{q},w_\psi) - L(\mathbf{q},w^c) \\ &= \mathbf{q}^T[\gamma(P^{\tilde{\pi}})^T - I](d_\psi - w^c). \end{aligned}$$

Rewriting the product of matrix and vectors into expectations,

$$L(\mathbf{q},w_\psi) = \mathbb{E}_{(x',a;)\sim d^\mu_{x,a}}\left[(w_\psi(x',a') - w(x',a'))(\Pi q)(x',a')\right],$$

where $(\Pi q)(x',a') \coloneqq \gamma\mathbb{E}_{a''\sim\tilde{\pi}(\cdot|x'')}\left[q(x'',a'')\right] - q(x',a')$ where $x'' \sim p(\cdot|x',a')$. Interestingly, here $(\Pi q)(x',a')$ could be interpreted as a reward such that if policy $\tilde{\pi}$ is executed, the Q-function would be $q(x',a')$. Following the techniques of [10], it is straightforward to show that when $q(x',a',x^*,a^*) = \sum_{t\geq 0}\gamma^t T^t_{\tilde{\pi}}(x',a'|x,a)$, we have $(\Pi q)(x',a') = \delta(x'=x^*,a'=a^*)$. Here, importantly, because $\tilde{\pi}$ is a sub-probability measure, $T^t_{\tilde{\pi}}$ exists and $\sum_{t\geq 0} T^t_{\tilde{\pi}}$ converges. As a result, with this choice of $\mathbf{q}(x^*,a^*)$, we have $L(\pm\mathbf{q}(x^*,a^*),w_\psi) = \pm(w_\psi(x^*,a^*) - w(x^*,a^*))$. Then it follows that when $\{\pm q(x',a',x^*,a^*),\forall(x,a)\} \in \mathcal{Q}$, the error $|w_\psi(x^*,a^*) - w(x^*,a^*)|$ is upper bounded by $\max_{\mathbf{q}\in\mathcal{Q}} L(\mathbf{q},w_\psi)$. □

Proposition 5.3. Assume that $\mathcal{Q}_b = \{\pm\delta(x=x^*,a=a^*),\forall(x^*,a^*)\} \subset \mathcal{Q}$. When $c_t = \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}$ and $w^c = w^{\pi,\mu}$, the contraction rate of $\mathcal{M}^{w_\psi}$ is upper bounded as $\eta^{w_\psi}_{x,a} \leq \max_{\mathbf{q}\in\mathcal{Q}} L(\mathbf{q},w_\psi)$.

Proof. Assume $\mathcal{Q}_b \subset \mathcal{Q}$. Based on Eqn (16), we deduce the following

$$\max_{\mathbf{q}\in\mathcal{Q}} L(\mathbf{q},w_\psi) = \max_{\mathbf{q}\in\mathcal{Q}} \mathbf{q}^T[(1-\gamma)\delta_{x,a} + \gamma(P^\pi)^T w_\psi - w_\psi] = \sum_{x',a'}\left|(1-\gamma)\delta_{x,a} + \gamma(P^\pi)^T w_\psi - w_\psi\right|(x',a') = \eta_{w_\psi}.$$

Here, the maximizer $\mathbf{q}^* \in \mathcal{Q}_b$ is

$$\mathbf{q}^*(x,a) = \text{sign}\left((1-\gamma)\delta_{x,a} + \gamma(P^\pi)^T w_\psi - w_\psi]\right),$$

where $\text{sign}(x)$ is the element-wise sign function. □

Proposition 4.3. The following holds for the sequence of values produced by relaxed LPs,

$$\left\|Q_{t+1} - Q^\pi\right\|_\infty \leq \eta\left\|Q_t - Q^\pi\right\|_\infty.$$

Proof. Let $Q_t(x,a)$ be the set of LP objectives at iteration t, and assign them to the objective coefficients of LPs at iteration $t+1$. This operation is defined through an equivalent operator $\mathcal{R}$. Recall that we abuse notations and denote $Q_{t+1}, Q_t \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ as vector Q-functions. Let d^* be the optimal solution to the $\text{LP}^{(t)}(x,a)$, then by construction

$$Q_{t+1}(x,a) = \delta^T_{x,a}Q_t + (1-\gamma)^{-1}(d^*)^T(R + \gamma P^\pi Q_t - Q_t).$$

Then recall that the constraints in Eqn (9) imply that

$$\left\|(1-\gamma)\delta_{x,a} + \gamma(P^\pi)^T d^* - d^*\right\|_1 \leq (1-\gamma)\eta.$$

This implies that the iteration $Q_{t+1} \leftarrow Q_t \equiv \delta_{x,a}^T Q_t + (1-\gamma)^{-1}(d^*)^T(R+\gamma P^\pi Q_t - Q_t)$ is contractive. In addition, the fixed point of this process is Q^π. We then conclude the desired result.

Discussion on more general results. The above proof relies on the important fact that the feasible set defined in $\text{LP}(x,a)$ in Eqn (9) corresponds to a set of d such that the iteration process is contractive. Hence, if we choose any arbitrary element $\tilde{d} \in \mathcal{D}_{x,a}$, and define

$$Q_{t+1} \leftarrow \delta_{x,a}^T Q_t + (1-\gamma)^{-1}(\tilde{d})^T(R+\gamma P^\pi Q_t - Q_t),$$

we still have the contraction $\left\|Q_{t+1} - Q^\pi\right\|_\infty \leq \eta \left\|Q_t - Q^\pi\right\|_\infty$.

□

Corollary 4.4. For any (x,a), let $w_{x,a}^* = \frac{d^*}{d_{x,a}^\mu} \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ and d^* is the optimal solution to $\text{LP}^{(t)}(x,a)$, then $\eta_{x,a}^{w_{x,a}^*} \leq \eta$ and

$$Q_{t+1}(x,a) = \mathcal{M}^{w_{x,a}^*} Q_t(x,a).$$

Proof. By construction, the following is true

$$\mathcal{M}^w Q = \delta_{x,a}^T Q_t + (1-\gamma)^{-1} d^T (R + \gamma P^\pi Q_t - Q_t),$$

where $d = w \odot d_{x,a}^\mu$. Recall also by construction,

$$Q_{t+1} \leftarrow \delta_{x,a}^T Q_t + (1-\gamma)^{-1}(d^*)^T(R+\gamma P^\pi Q_t - Q_t).$$

Combining the above two equations directly implies the desired result. □

B Relations between different stochastic estimators

$$\begin{tikzcd}
\hat{\mathcal{R}}_\tau^c \arrow[r] & \hat{\mathcal{R}}^c \\
\hat{\mathcal{M}}_\tau^{w_c} \arrow[u] \arrow[r] & \hat{\mathcal{M}}^{w_c}
\end{tikzcd}$$

Figure 4: Visualization of relations between stochastic estimates to different operators. Blue arrows represent marginalization over the random time variables τ; the red arrow represents marginalization over the random state-action pair (x_τ, a_τ). Under suitable conditions, the directions of the arrows indicate potential variance reductions.

In Figure 4, we show relations between stochastic estimates to different operators. Blue arrows represent marginalization over the random time variables τ; the red arrow represents marginalization over the random state-action pair (x_τ, a_τ). Under suitable conditions, the directions of the arrows indicate potential variance reductions. We see that under suitable conditions outlined in Proposition 4.1, we have $\mathbb{V}[\mathcal{M}_\tau^{w^c}] \geq \mathbb{V}[\mathcal{M}^{w^c}]$, $\mathbb{V}[\mathcal{R}_\tau^c] \geq \mathbb{V}[\mathcal{R}^c]$ and $\mathbb{V}[\mathcal{R}_\tau^c] \geq \mathbb{V}[\mathcal{M}_\tau^{w^c}]$. However, it is not clear what is the ordering of the variance between $\mathcal{R}^c$ and $\mathcal{M}^{w^c}$.

C Marginalized V-trace operator

The V-trace operator [24] is defined for value functions $V \in \mathbb{R}^{|\mathcal{X}|}$. Given a target policy π and a behavior policy μ, the operator $\mathcal{R}^{c,\rho}$ is parameterized by step-wise trace coefficients $c(x,a)$ and $\rho(x,a)$. In particular,

$$\mathcal{R}^{c,\rho} V(x) \coloneqq V(x) + \mathbb{E}_{\mu}\left[\sum_{t\geq 0} \gamma^t (c_0 ... c_{t-1}) \rho_t \Delta_t \;\middle|\; x_0 = x\right], \tag{17}$$

where $\Delta_t = \bar{r}_t + \gamma \mathbb{E}_{x' \sim p(\cdot|x_t,a), a\sim\mu(\cdot|x_t)}\left[V(x_{t'})\right] - V(x_t)$ is the TD-error at step t. Here, ρ_t determines the fixed point of the operator, while ρ_t, c_t jointly determine the contraction rate. Consider defining a marginalized V-trace operator $\mathcal{M}^{w,\rho} V(x_0)$ as below

$$\mathcal{M}^{w,\rho} V(x) \coloneqq V(x) + (1-\gamma)^{-1} \mathbb{E}_{x' \sim d^{\mu}_{x,a}}\left[w_x(x') \rho(x',a') \Delta(x',a')\right], \tag{18}$$

where $d^{\mu}_x(x) \coloneqq (1-\gamma)\sum_{t\geq 0} P_{\mu}(x_t = x' | x_0 = x)$ is the discounted visitation distribution under μ. here, $w(x')$ is state-dependent, $\rho(x',a')$ is state-action dependent and $\Delta(x',a') \coloneqq \bar{r}(x',a') + \gamma \mathbb{E}_{x'' \sim p(\cdot|x',a'), a' \sim \mu(\cdot|x')}\left[V(x'')\right] - V(x')$.

C.1 State-marginalized V-trace operator.

Consider setting $\rho(x,a)$ and the V-trace step-wise traces. Define TD weights $w^c_x(x')$, which is computed as

$$w^c_x(x') \coloneqq \mathbb{E}_{\mu}\left[\sum_{t\geq 0} \gamma^t (c_0 ... c_{t-1}) \mathbb{I}[x_t = x'] \;\middle|\; x_0 = x\right]. \tag{19}$$

It is then straightforward to show that the multi-step operator and the marginalized operator are equivalent in expectation $\mathcal{R}^{c,\rho} \equiv \mathcal{M}^{w,\rho}$. The trace coefficient is obtained via conditional expectation as follows.

Proposition C.1. Let τ be an integer-valued random time, such that $P(\tau = n) = (1-\gamma)\gamma^n, \forall n \geq 0$. For V-trace, given any step-wise trace coefficient c_t, its equivalent TD weights $w(x')$ is

$$w^c_x(x') = \mathbb{E}_{\mu,\tau}\left[\left(\Pi_{1\leq s\leq \tau-1} c_s\right) \mid x_\tau = x', x_0 = x\right].$$

Now define $d^{w^c}_x(x') \coloneqq w^c_x(x') d^{\mu}_x(x')$. It can be shown that $d^{w^c}_x(x') \in \mathbb{R}^{|\mathcal{X}|}$ also satisfy fixed point equations.

Proposition C.2. The following Bellman equation holds for the step-wise trace coefficient c_t and $d^{w^c}_{x,a}(x')$

$$d^{w^c}_x(x') = (1-\gamma)\delta(x' = x) + \gamma \sum_{x',a'} d^{w^c}_x(x') c(x',a') \mu(a'|x') p(x''|x',a'),$$

where δ is the Dirac function. Let $P^{\tilde{\pi}} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ be a transition matrix such that $P^{\tilde{\pi}}(x,a) = \sum_{a'} p(y|x',a')\tilde{\pi}(a'|x') = \sum_a p(y|x',a')\mu(a'|x')c(x',a')$. Then equivalently, in matrix form,

$$d^{w^c}_x = (1-\gamma)\delta_x + \gamma (P^{\tilde{\pi}})^T d^{w^c}_x. \tag{20}$$

Based on the Bellman equation in Eqn (20), it is possible to estimate $w^c(x)$ from the behavior data under μ. In particular, given a starting state pair x, let $w_\psi(x') \approx d^{w^c}_x(x') \in \mathbb{R}^{|\mathcal{X}|}$ be a parameteric function used for estimating $d^{w^c}_x(x')$. With a critic function $\mathbf{q} \in \mathcal{Q} \subset \mathbb{R}^{\mathcal{X}}$, formulate the objective

$$\begin{aligned} L(\mathbf{q}, w_\psi) &\coloneqq \mathbf{q}^T\left[(1-\gamma)\delta_x + \gamma (P^{\tilde{\pi}})^T w_\psi - w_\psi\right] \\ &= (1-\gamma) q(x) + \mathbb{E}_{(y)\sim d^{\mu}_a}\left[\Delta(y)\right]. \end{aligned} \tag{21}$$

Here, the TD-error $\Delta(y) \coloneqq \left(\mathbb{E}_{b\sim\pi(\cdot|y), y'\sim p(\cdot|y,b)}\left[q(y')c(y,b)\right] - q(y)\right) w_\psi(y)$. By solving a saddle point optimization problem of the above objective $\min_{w_\psi} \max_{\mathbf{q}} L(\mathbf{q}, w_\psi)$, we find $w_\psi \approx d^{w^c}_x$.

D Multi-step RL algorithms

Motivated by previous theoretical insights, we seek a practical algorithm which could combine the benefits of multi-step TD-learning and estimations of TD weights.

D.1 Multi-step RL algorithms with TD weights

We focus on the actor-critic setup where the algorithm maintains a target policy π and a Q-function critic $Q_\theta(x,a)$. Here, the policy π could be either parameterized $\pi = \pi_\phi$ or defined by the Q-function, e.g. the greedy policy. The algorithm collects data with behavior policy μ.

To estimate TD weights, we parameterize the scoring function $q_\eta(x,a;x_0,a_0)$ and estimator $w_\psi(x,a;x_0,a_0)$, both taking as inputs the starting state-action pair (x_0,a_0) and the target pair (x,a). For simplicity of notations, we omit the dependency on (x_0,a_0). Given a trajectory $(x_t,a_t,r_t)_{t=0}^\infty \sim \mu$, we approximate the loss function in Eqn (10) with stochastic samples,

$$\hat{L}(\eta,\psi) \coloneqq (1-\gamma)q_\eta(x_0,a_0) + (1-\gamma)\sum_{t=0}^\infty \gamma^t \hat{\Delta}_t, \tag{22}$$

where $\hat{\Delta}_t = \gamma \mathbb{E}_\pi \left[q_\eta(x_{t+1},\cdot)w_\psi(x_{t+1},\cdot)\right] - q_\eta(x_t,a_t)w_\psi(x_t,a_t)$. Following prior methods on scaling saddle-point optimization to neural networks (e.g. [11]), we approximate the optimal solution to Eqn (10) via stochastic gradient descents (ascents) on the empirical loss $\eta \leftarrow \eta + \alpha\nabla_\eta \hat{L}(\eta,\psi), \psi \leftarrow \psi - \alpha \nabla_\psi \hat{L}(\eta,\psi)$.

At policy evaluation stage, we construct the Q-function targets with $\mathcal{M}^w Q(x_0,a_0)$. From the trajectory $(x_t,a_t,r_t)_{t=0}^\infty$, compute the following Q-function target

$$\mathcal{M}^w Q(x_0,a_0) \coloneqq Q_{\theta^-}(x_0,a_0) + (1-\gamma)^{-1}\sum_{t=0}^\infty \gamma^t w_\psi(x_t,a_t)\hat{\Delta}_t, \tag{23}$$

where $\hat{\Delta}_t = r_t + \gamma\mathbb{E}_\pi\left[Q_{\theta^-}(x_{t+1},\cdot)\right] - Q_{\theta-}(x_t,a_t)$. Then the Q-function is optimized by minimizing $(Q_\theta(x_0,a_0) - \mathcal{M}^w(x_0,a_0))^2$.

Algorithm 1 Multi-step RL with TD weights

```
Require: policy $\pi$, Q-function critic $Q_\theta(x,a)$, density estimator $w_\psi(x)$, critic $q_\eta$ and learning rate $\alpha \geq 0$
  while not converged do
    1. Collect data $(x_t,a_t,r_t)_{t=0}^\infty \sim \mu$ and save to the buffer $\mathcal{D}$
    2. Construct the empirical loss for marginalized estimation based on Eqn (22). Optimize $\eta, \psi$ by alternating gradient descents (ascents): $\eta \leftarrow \eta + \alpha\nabla_\eta\hat{L}(\eta,\psi), \psi \leftarrow \psi - \alpha\nabla_\psi\hat{L}(\eta,\psi)$.
    3. Construct Q-function targets based on Eqn (23). Optimize $\theta$: $\theta \leftarrow \theta - \alpha\nabla_\theta(Q_\theta - \mathcal{M}^w Q(x_0,a_0))^2$.
    4. Improve the policy by either policy gradient or being greedy with respect to the new Q-function $Q_\theta(x,a)$.
    5. Update target network $\theta^- \leftarrow \theta$.
  end while
```

E Fenchel-duality based approach to estimating TD weights

In this section, we introduce Fenchel-duality based approaches to off-policy evaluation [11, 33, 13]. While different in details, a common feature of this family of work is to convert the off-policy evaluation problem into a convex-concave optimization problem. Here, we focus on the initial formulation *Dualdice*. We start by introducing this algorithm and then discuss how to extend this framework to estimate TD weights.

E.1 Background on Dualdice

Following [11], consider the following optimization problem with argument $w \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$

$$J(w) = \frac{1}{2}\mathbb{E}_{(x',a')\sim d^\mu_{x,a}}\left[w^2(x',a')\right] - \mathbb{E}_{(x',a')\sim d^\pi_{x,a}}\left[w(x',a')\right].$$

This objective is minimized at $w(x',a') = \frac{d^\mu_{x,a}(x',a')}{d^\pi_{x,a}(x',a')}$. Note that the original derivation from [11] focuses on the discounted visitation distribution $d^\mu_x(x')$) without conditioning on the initial action a_0 as in our case $d^\mu_{x,a}(x',a')$, because they focus on policy evaluation of a single starting state x. However, it is straightforward to extend their results. By Fenchel duality, one could further show that the above optimization could be transformed into the following saddle-point problem with $v, \psi \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$,

$$\min_v \max_\psi \ \mathbb{E}_{x',a'\sim d^\mu_{x,a}, x''\sim p(\cdot|x,a), a''\sim\pi(\cdot|x')}\left[(v(x',a') - \gamma v(x'',a''))\psi(x',a') - \frac{\psi^2(x',a')}{2}\right] - (1-\gamma)v(x,a). \tag{24}$$

The main motivation for proposing the saddle-point optimization problem is to bypass the double sampling issue [34]. The saddle point of Eqn (24) is (v^*, ψ^*) and $\psi^*(x', a') = \frac{d^{\pi}_{x,a}(x',a')}{d^{\mu}_{x,a}(x',a')}$. See [11] for details

E.2 Fenchel duality-based estimation for TD weights

Now we introduce the extension to TD weights. Given a step-wise trace coefficient $c(x, a)$ and its equivalent TD weights $w^c(x, a)$, recall that we define $d^{w^c}_{x,a}(x', a') \coloneqq w^c(x', a') \cdot d^{\mu}_{x,a}(x', a')$. Consider the following objective, whose optimal solution is $w^c_{x,a}(x', a')$.

$$\arg\min_{w} J(w) = \frac{1}{2}\mathbb{E}_{(x',a')\sim d^{\mu}_{x,a}}\left[w^2(x',a')\right] - \mathbb{E}_{(x',a')\sim d^{w^c}_{x,a}}\left[w(x',a')\right]. \tag{25}$$

First, we define $\tilde{\pi}(a|x) \coloneqq c(x, a)\mu(a|x)$. We assume that $c(x, a)$ is such that $\tilde{\pi}(\cdot|x)$ is a sub-probability vector. this is satisfied in the context of general Retrace ($c(x, a) \leq \frac{\pi(a|x)}{\mu(a|x)}$ [6]).

Now, define variables $v(x', a')$ such that $v(x', a') = w(x', a') + \gamma\mathbb{E}_{x''\sim p(\cdot|x',a'),a''\sim\tilde{\pi}(\cdot|x'')}\left[v(x'', a'')\right]$. Note that such a quantity $v(x', a')$ exists and is unique. To see why, it is straightforward to verify that $v(x', a')$ is the fixed point of the operator $\mathcal{T}^{\tilde{\pi}}$, defined as $\mathcal{T}^{\tilde{\pi}}Q(x', a') \coloneqq w(x', a') + \mathbb{E}_{x''\sim p(\cdot|x',a'),a''\sim\tilde{\pi}(\cdot|x'')}\left[Q(x'', a'')\right]$. Because $\gamma < 1$ and $\tilde{\pi}(\cdot|x'')$ is a sub-probability vector, this operator is contractive and has a unique fixed point. As a result, starting from $w(x', a')$, by applying $(\mathcal{T}^{\tilde{\pi}})^k w(x', a')$ and let $k \to \infty$ we obtain $v(x', a')$. In vector notations, the second term of Eqn (25) writes

$$(d^{w^c}_{x,a})^T w = (d^{w^c}_{x,a})^T(\mathbf{v} - \gamma P^{\tilde{\pi}}\mathbf{v}) = (d^{w^c}_{x,a} - \gamma(P^{\tilde{\pi}})^T d^{w^c}_{x,a})^T\mathbf{v} = (1-\gamma)\delta^T_{x,a}v = (1-\gamma)v(x,a),$$

where the second to last equality stems from the Bellman equation of $d^{w^c}_{x,a}$ in Eqn (15).

The integrand of the first term can be rewritten as $\left((v - \gamma P^{\tilde{\pi}}v)(x', a')\right)^2$, but directly plugging in the transition matrix $P^{\tilde{\pi}}$ results in the double-sampling problem [34]. To bypass this, we follow the exact same procedure as [11] and propose the following saddle-point optimization problem.

$$\min_{v}\max_{\psi}\ \mathbb{E}_{x',a'\sim d^{\mu}_{x,a},x''\sim p(\cdot|x',a'),a''\sim\tilde{\pi}(\cdot|x'')}\left[(v(x',a') - \gamma v(x'',a''))\psi(x',a') - \frac{\psi^2(x',a')}{2}\right] - (1-\gamma)v(x,a). \tag{26}$$

The saddle point solution (v^*, ψ^*) will be such that $\psi^*(x', a') = w^c_{x,a}(x', a')$. Note that the only difference between Eqn (26) and Eqn (24) is the target policy. Alternatively, one could interpret the new objective in Eqn (26) as executing the original dualdice algorithm but with the behavior policy $\tilde{\pi}$, which is in general a sub-probability policy.

F Experiment

F.1 Details on tabular estimations of TD weights

We adopt tabular representations for w_ψ for both the chain MDP and Open World MDP. For tabular MDPs with $|\mathcal{X}|$ states and $|\mathcal{A}|$ actions, we represent w_ψ as a $|\mathcal{X}||\mathcal{A}| \times |\mathcal{X}||\mathcal{A}|$ matrix. When both the critic $\mathbf{q} \in \mathcal{Q}$ and the estimates w_ψ are tabular represented, there is no need for solving the saddle point optimization. In fact, one can directly derive solutions to the estimates given off-policy samples. We summarize the algorithmic procedure for estimating TD weights in Algorithm 2.

Given a trajectory $(x_t, a_t, r_t)_{t=0}^{\infty}$, Algorithm 2 specifies how to construct empirical estimates $\hat{w}$ and update the table w_{x_0,a_0}, i.e., the TD weights with initial state (x_0, a_0). However, all state-action pairs along the trajectory could be seen as initial states. To get updates for all such pairs, we need to loop through initial pairs along the trajectory.

Remarks. We can interpret Algorithm 2 as a direct implementation of the Monte-Carlo estimation to the TD weights as defined in Eqn (6). This bears close resemblance to marginalized estimation techniques adopted in [23].

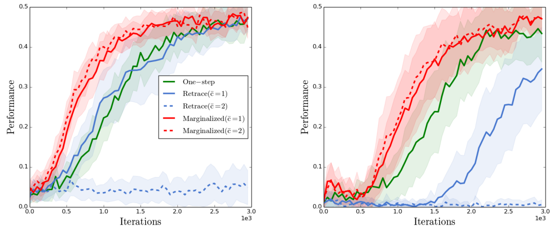
F.2 Additional experiment results

F.2.1 Chain MDP

Details on Q-function estimation. At each iteration t, the agent collects $N = 1$ trajectory $(x_t, a_t, r_t)_{t=0}^{\infty}$. The agent maintains a Q-function $Q^{(t)}(x, a)$. Along the trajectory, we use an operator baseline to generate

```
Algorithm 2 Tabular estimation of TD weights
Require: Table $w$ of size $|\mathcal{X}||\mathcal{A}| \times |\mathcal{X}||\mathcal{A}|$ initialized with zeros
  while not converged do
    1. Collect a trajectory $(x_t, a_t, r_t)_{t=0}^{\infty} \sim \mu$
    2. Construct cumulative step-wise traces along the trajectory: define $\hat{C}(x_t, a_t) := (1-\gamma)\gamma^t \left(\Pi_{1 \leq s \leq t} c(x_s, a_s)\right)$ for all $t \geq 0$.
    3. Accumulate cumulative step-wise traces per state-action:
       $$\hat{w}(x,a) := \frac{\sum_{t \geq 0} \hat{C}(x_t, a_t)\mathbb{I}[x_t = x', a_t = a']}{\sum_{t \geq 0} \mathbb{I}[x_t = x, a_t = a]}, \forall (x,a).$$
    If the denominator is zero, define the ratio to be zero.
    4. Update the estimate $\hat{w}(x,a)$ and set $w_{x_0,a_0}(x,a) \leftarrow (1-\alpha) w_{x_0,a_0}(x,a) + \alpha \hat{w}(x,a)$ for all $(x,a)$ with $\alpha = 0.1$.
  end while
```

estimates $\hat{Q}(x_t, a_t)$. Then the Q-function is updated as $Q^{(t+1)}(x_t, a_t) \leftarrow (1-\alpha)Q^{(t)}(x_t, a_t) + \alpha \hat{Q}(x_t, a_t)$ with $\alpha = 0.1$. The relative errors in Figure 4 are computed as $\sum_a \frac{|Q^{(t)}(x_0,a) - Q^{\pi}(x_0,a)|}{Q^{\pi}(x_0,a)}$, i.e., an average measure of prediction error at the initial state x_0 (the leftmost state of the chain). Here, $Q^{\pi}(x,a)$ is computed analytically from the MDP.



(a) Soft PI $\alpha = 0.1$

(b) *Hard* PI $\alpha = 1$

Figure 5: Comparison of RL algorithms based on baseline operators. Each plot is averaged over 50 runs. The x-axis shows the number of iterations and y-axis shows the performance of algorithms.

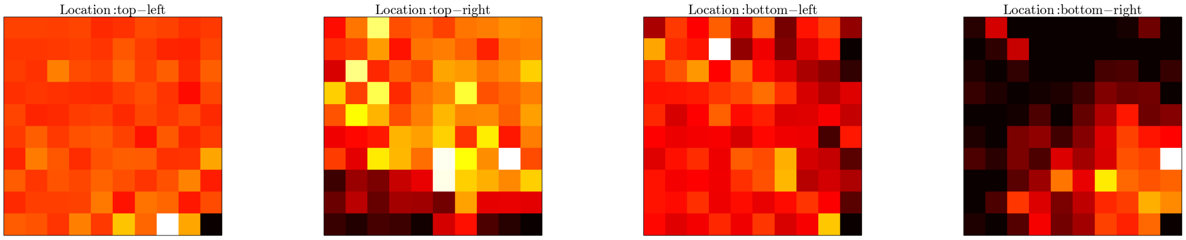
F.2.2 Open world

Visualization of TD weights. In Figure 6, we visualize the TD weights learned by tabular representations. Recall that in general, $w_\psi \approx w^c$ is a matrix – it takes two pairs of state-action, (x,a) and (x',a'). Here (x,a) is the initial state-action pair while (x',a') is the typical argument. In the four subplots of Figure 6, we each fix the initial location (x,a) and visualize TD weights as a function of (x',a') as heat maps.

Overall, we see that the learned TD weights reflect the intuition of correct credit assignment. In Figure 6(d), where the initial state is located near the terminal state (bottom-right), it assigns low weights to most state-action pairs except near the bottom-right corner. In this case, the intuition is that Bellman errors at state-action pairs far from the bottom right should contribute much less to the estimation on average, because the random policy μ has a small chance of visiting them.

Results on policy optimization. We also consider the setup of a full off-policy optimization algorithm: policy iteration (PI): the behavior policy μ is always uniformly random, the target policy $\pi^{(0)}$ is initialized as random. At iteration i, the new policy is computed as $\pi^{(i+1)} = (1-\alpha)\pi^{(i)} + \alpha\pi_{\text{target}}$ where π_{target} is the greedy policy with respect to the Q-function estimate $\hat{Q}$ at iteration i. In Figure 5(a), we carry out *soft* PI by setting $\alpha = 0.1$; and in Figure 5(b), hard PI by setting $\alpha = 1$.

To evaluate the performance, we compare the average returns starting from uniformly random sampled states, estimated via MC estimates. For the marginalized operators, the performance gains in the *one-shot* off-policy evaluation seem to carry over to the downstream optimization; however, this is not the case for Retrace. where it



(a) Top left

(b) Top right

(c) Bottom left

(d) Bottom right

Figure 6: Visualization of TD weights in the open world problem. All four plots show the trace estimation with different starting state, located at the top left, top right, bottom left and bottom right of the square. The trace for a state is the average of the TD weights for all actions at the state. Recall that given an initial state, the TD weights are a function of future states, which spans the entire state space.

obtains a similar performance as the one-step operator for the soft PI; for the hard PI, because $\pi^{(i)}, i \geq 1$ are greedy policies, it is likely to cut traces quickly. In this case, Retrace does not seem to retain advantages over the one-step operator and slow down the optimization.

F.3 Further details on deep RL experiments

Benchmarks. For the deep RL implementations of the algorithms, we focus on continuous control tasks [29, 35], with various simulation engines, such as MuJoCo [36] and Bullet physics [37]. These benchmarks generally consist of locomotion tasks defined with robotics systems, with state space $\mathcal{X}$ the sensory inputs such as velocities and joints, and $\mathcal{A}$ the position or toeque controls. See documentations such as [35] for details. In our experiments, we use (D) to stand for DeepMind control suite [35] and (B) to stand for bullet physics [37].

Algorithms. We consider twin-delayed deep deterministic policy gradient (TD3) [28] as the baseline algorithm. By default, the algorithm maintains a deterministic policy $\pi_\theta(x)$ and Q-function critic $Q_\theta(x,a)$. The policy is updated by the gradients $\nabla_\theta Q_\phi(x, \pi_\theta(x))$. The critic is updated by regression against Q-function targets, such that $Q_\phi \approx Q^\pi$. Different algorithms vary in ohw the Q-function targets are defined. In general, they are defined by stochastic estimates of the evaluation operator $\mathcal{R}Q(x,a)$. For example, the vanilla TD3 constructs the target as the one-step target $Q_{\text{target}}(x,a) = r(x,a) + \gamma Q_\phi(x', \pi_\theta(x'))$. TD3 also introduces a set of techniques, such as double Q-learning [38, 39] and target networks [1] to stabilize updates.

Per Algorithm 1, marginalized operators also need a density estimator w_ψ and a discriminator q_η. They are trained via the objective defined in Eqn (10),

$$\min_\psi \max_\eta L(q_\eta, w_\psi) = (1-\gamma) q(x,a) + \mathbb{E}_{(x',a') \sim d^\mu_{x,a}, x'' \sim p(\cdot|x',a')} \left[\delta(x', a', x'')\right].,$$

where data $(x',a') \sim d^\mu_{x,a}, x'' \sim p(\cdot|x',a')$ are equivalently sampled as tuples (x', a', x'') from the replay buffer. We can construct $Q_{\text{target}}(x,a) = \mathcal{M}^{w_\psi} Q(x,a)$. Parameters ψ and η are optimized with alternating gradient descents (ascents). See Appendix D for further algorithmic details.

Baseline multi-step algorithm. We implement a variant of Retrace [6] as the baseline multi-step algorithm. Such algorithms start with a trajectory $(x_t, r_t, a_t)_{t=0}^\infty$ starting from (x,a) such that $x_0 = x, a_0 = a$, Q-function targets are computed recursively

$$\hat{Q}_i = r_i + \gamma Q_{\phi^-}(x_{i+1}, \pi_{\theta^-}(x_{i+1})) + \gamma c_i \left(\hat{Q}_{i+1} - \tilde{Q}_{i+1}\right). \tag{27}$$

Here, parameters ϕ^-, θ^- are delayed copies of the parameters ϕ, θ [1]. The coefficients $c(x,a) = \lambda \cdot \min\{1, \frac{\pi(a|x)}{\mu(a|x)}\}$. By Retrace, the Q-function $\tilde{Q}_{i+1} = Q(x_{i+1}, a_{i+1})$, which we find to not work stably in practice. Instead, we use $\tilde{Q}_{i+1} = Q_{\phi^-}(x_{i+1}, \pi_{\theta^-}(x_{i+1}))$. Throughout the experiments, we use $\lambda = 0.7$ for the multi-step algorithms.

Implementation details and other hyper-parameters. All implementations are built on SpinningUp [40]. Please refer to the code base for all missing details on network architecture and hyper-parameters.

Architecture and hyper-parameters. All policy networks π_θ, Q-function networks Q_ϕ, discirminator q_η and estimator w_ψ share the same torso networks. After the input layer, they have 2 layers of hidden units each of size 256. The inputs to the policy network π_θ are only the state variables x, while for all other networks are the concatenated state-action variables $[x, a]$. The discriminator output is squashed between $[-1, 1]$ via $\tanh(x)$ activation; the estimator w_ψ output is transformed by $f(x) = \log(1 + \exp(x))$ to ensure that it is strictly non-negative. Finally, the density estimator w_ψ is transformed across batch $\tilde{w}(x_i, a_i) = \frac{(w(x_i, a_i))^T}{\sum_j (w(x_j, a_j))^T}$ to ensure stability, where $T = 0.1$.

All networks are trained with sub-sampling of mini-batches from a replay buffer. Each mini-batch is of size 100. All networks are trained with Adam [41] optimizers with learning rates 10^{-3} except for the estimator, where the learning rate is 10^{-4}.