DualDICE: Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections

Ofir Nachum* **Yinlam Chow*** **Bo Dai** **Lihong Li**
Google Research
{ofirnachum,yinlamchow,bodai,lihong}@google.com

Abstract

In many real-world reinforcement learning applications, access to the environment is limited to a fixed dataset, instead of direct (online) interaction with the environment. When using this data for either evaluation or training of a new policy, accurate estimates of *discounted stationary distribution ratios* — correction terms which quantify the likelihood that the new policy will experience a certain state-action pair normalized by the probability with which the state-action pair appears in the dataset — can improve accuracy and performance. In this work, we propose an algorithm, DualDICE, for estimating these quantities. In contrast to previous approaches, our algorithm is agnostic to knowledge of the behavior policy (or policies) used to generate the dataset. Furthermore, it eschews any direct use of importance weights, thus avoiding potential optimization instabilities endemic of previous methods. In addition to providing theoretical guarantees, we present an empirical study of our algorithm applied to off-policy policy evaluation and find that our algorithm significantly improves accuracy compared to existing techniques.[1]

1 Introduction

Reinforcement learning (RL) has recently demonstrated a number of successes in various domains, such as games [25], robotics [1], and conversational systems [11, 18]. These successes have often hinged on the use of simulators to provide large amounts of experience necessary for RL algorithms. While this is reasonable in game environments, where the game is often a simulator itself, and some simple real-world tasks can be simulated to an accurate enough degree, in general one does not have such direct or easy access to the environment. Furthermore, in many real-world domains such as medicine [26], recommendation [19], and education [24], the deployment of a new policy, even just for the sake of performance evaluation, may be expensive and risky. In these applications, access to the environment is usually in the form of *off-policy* data [39], logged experience collected by potentially multiple and possibly unknown *behavior* policies.

State-of-the-art methods which consider this more realistic setting — either for policy evaluation or policy improvement — often rely on estimating (*discounted*) *stationary distribution ratios* or *corrections*. For each state and action in the environment, these quantities measure the likelihood that one's current *target* policy will experience the state-action pair normalized by the probability with which the state-action pair appears in the off-policy data. Proper estimation of these ratios can improve the accuracy of policy evaluation [21] and the stability of policy learning [12, 14, 22, 40]. In general, these ratios are difficult to compute, let alone estimate, as they rely not only on the probability that the target policy will take the desired action at the relevant state, but also on the probability that the target policy's interactions with the environment dynamics will lead it to the relevant state.

*Equal contribution.

[1] Find code at https://github.com/google-research/google-research/tree/master/dual_dice.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Several methods to estimate these ratios have been proposed recently [12, 14, 21], all based on the steady-state property of stationary distributions of Markov processes [15]. This property may be expressed locally with respect to state-action-next-state tuples, and is therefore amenable to stochastic optimization algorithms. However, these methods possess several issues when applied in practice: First, these methods require knowledge of the probability distribution used for each sampled action appearing in the off-policy data. In practice, these probabilities are usually not known and difficult to estimate, especially in the case of multiple, non-Markovian behavior policies. Second, the loss functions of these algorithms involve per-step importance ratios (the ratio of action sample probability with respect to the target policy versus the behavior policy). Depending on how far the behavior policy is from the target policy, these quantities may have large variance, and thus have a detrimental effect on stochastic optimization algorithms.

In this work, we propose *Dual stationary DIstribution Correction Estimation* (*DualDICE*), a new method for estimating discounted stationary distribution ratios. It is agnostic to the number or type of behavior policies used for collecting the off-policy data. Moreover, the objective function of our algorithm does not involve any per-step importance ratios, and so our solution is less likely to be affected by their high variance. We provide theoretical guarantees on the convergence of our algorithm and evaluate it on a number of off-policy policy evaluation benchmarks. We find that DualDICE can consistently, and often significantly, improve performance compared to previous algorithms for estimating stationary distribution ratios.

2 Background

We consider a Markov Decision Process (MDP) setting [32], in which the environment is specified by a tuple $\mathcal{M} = \langle S, A, R, T, \beta \rangle$, consisting of a state space, an action space, a reward function, a transition probability function, and an initial state distribution. A policy π interacts with the environment iteratively, starting with an initial state $s_0 \sim \beta$. At step $t = 0, 1, \cdots$, the policy produces a distribution $\pi(\cdot|s_t)$ over the actions A, from which an action a_t is sampled and applied to the environment. The environment stochastically produces a scalar reward $r_t \sim R(s_t, a_t)$ and a next state $s_{t+1} \sim T(s_t, a_t)$. In this work, we consider infinite-horizon environments and the γ-discounted reward criterion for $\gamma \in [0, 1)$. It is clear that any finite-horizon environment may be interpreted as infinite-horizon by considering an augmented state space with an extra terminal state which continually loops onto itself with zero reward.

2.1 Off-Policy Policy Evaluation

Given a *target* policy π, we are interested in estimating its value, defined as the normalized expected per-step reward obtained by following the policy:

$$\rho(\pi) := (1-\gamma) \cdot \mathbb{E}\big[\textstyle\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 \sim \beta, \forall t, a_t \sim \pi(s_t), r_t \sim R(s_t, a_t), s_{t+1} \sim T(s_t, a_t) \big]. \quad (1)$$

The off-policy policy evaluation (OPE) problem studied here is to estimate $\rho(\pi)$ using a fixed set $\mathcal{D}$ of transitions (s, a, r, s') sampled in a certain way. This is a very general scenario: $\mathcal{D}$ can be collected by a single behavior policy (as in most previous work), multiple behavior policies, or an oracle sampler, among others. In the special case where $\mathcal{D}$ contains entire trajectories collected by a known behavior policy μ, one may use *importance sampling* (IS) to estimate $\rho(\pi)$. Specifically, given a finite-length trajectory $\tau = (s_0, a_0, r_0, \ldots, s_H)$ collected by μ, the IS estimate of ρ based on τ is estimated by [31]: $(1-\gamma) \left(\prod_{t=0}^{H-1} \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} \right) \left(\sum_{t=0}^{H-1} \gamma^t r_t \right)$. Although many improvements exist [e.g., 9, 16, 31, 43], importance-weighting the entire trajectory can suffer from exponentially high variance, which is known as "the curse of horizon" [20, 21].

To avoid exponential dependence on trajectory length, one may weight the states by their *long-term* occupancy measure. First, observe that the policy value may be re-expressed as,

$$\rho(\pi) = \mathbb{E}_{(s,a) \sim d^\pi, r \sim R(s,a)}[r],$$

where

$$d^\pi(s, a) := (1-\gamma) \textstyle\sum_{t=0}^{\infty} \gamma^t \Pr\left(s_t = s, a_t = a \mid s_0 \sim \beta, \forall t, a_t \sim \pi(s_t), s_{t+1} \sim T(s_t, a_t)\right), \quad (2)$$

is the *normalized discounted stationary distribution* over state-actions with respect to π. One may define the discounted stationary distribution over states analogously, and we slightly abuse notation

by denoting it as $d^\pi(s)$; note that $d^\pi(s,a) = d^\pi(s)\pi(a|s)$. If $\mathcal{D}$ consists of trajectories collected by a behavior policy μ, then the policy value may be estimated as,

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\mu, r\sim R(s,a)}\left[w_{\pi/\mu}(s,a)\cdot r\right] ,$$

where $w_{\pi/\mu}(s,a) = \frac{d^\pi(s,a)}{d^\mu(s,a)}$ is the *discounted stationary distribution correction*. The key challenge is in estimating these correction terms using data drawn from d^μ.

2.2 Learning Stationary Distribution Corrections

We provide a brief summary of previous methods for estimating the stationary distribution corrections. The ones that are most relevant to our work are a suite of recent techniques [12, 14, 21], which are all essentially based on the following steady-state property of stationary Markov processes:

$$d^\pi(s') = (1-\gamma)\beta(s') + \gamma\sum_{s\in S}\sum_{a\in A} d^\pi(s)\pi(a|s)T(s'|s,a),\ \ \forall s'\in S, \tag{3}$$

where we have simplified the identity by restricting to discrete state and action spaces. This identity simply reflects the conservation of flow of the stationary distribution: At each timestep, the flow out of s' (the LHS) must equal the flow into s' (the RHS). Given a behavior policy μ, equation 3 can be equivalently rewritten in terms of the stationary distribution corrections, i.e., for any given $s'\in S$,

$$\mathbb{E}_{(s_t,a_t,s_{t+1})\sim d^\mu}\left[\text{TD}(s_t,a_t,s_{t+1}\mid w_{\pi/\mu})\ \middle|\ s_{t+1}=s'\right] = 0\,, \tag{4}$$

where

$$\text{TD}(s,a,s'\mid w_{\pi/\mu}) := -w_{\pi/\mu}(s') + (1-\gamma)\beta(s') + \gamma w_{\pi/\mu}(s)\cdot\frac{\pi(a|s)}{\mu(a|s)}\,,$$

provided that $\mu(a|s) > 0$ whenever $\pi(a|s) > 0$. The quantity TD can be viewed as a *temporal difference* associated with $w_{\pi/\mu}$. Accordingly, previous works optimize loss functions which minimize this TD error using samples from d^μ. We emphasize that although $w_{\pi/\mu}$ is associated with a temporal difference, it does not satisfy a Bellman recurrence in the usual sense [2]. Indeed, note that equation 3 is written "backwards": The occupancy measure of a state s' is written as a (discounted) function of *previous* states, as opposed to vice-versa. This will serve as a key differentiator between our algorithm and these previous methods.

2.3 Off-Policy Estimation with Multiple Unknown Behavior Policies

While the previous algorithms are promising, they have several limitations when applied in practice:

- The off-policy experience distribution d^μ is with respect to a single, Markovian behavior policy μ, and this policy must be known during optimization.[2] In practice, off-policy data often comes from multiple, unknown behavior policies.
- Computing the TD error in equation 4 requires the use of per-step importance ratios $\pi(a_t|s_t)/\mu(a_t|s_t)$ at every state-action sample (s_t,a_t). Depending on how far the behavior policy is from the target policy, these quantities may have high variance, which can have a detrimental effect on the convergence of any stochastic optimization algorithm that is used to estimate $w_{\pi/\mu}$.

The method we derive below will be free of the aforementioned issues, avoiding unnecessary requirements on the form of the off-policy data collection as well as explicit uses of importance ratios. Rather, we consider the general setting where $\mathcal{D}$ consists of *transitions* sampled in an unknown fashion. Since $\mathcal{D}$ contains rewards and next states, we will often slightly abuse notation and write not only $(s,a)\sim d^\mathcal{D}$ but also $(s,a,r)\sim d^\mathcal{D}$ and $(s,a,s')\sim d^\mathcal{D}$, where the notation $d^\mathcal{D}$ emphasizes that, unlike previously, $\mathcal{D}$ is not the result of a single, known behavior policy. The target policy's value can be equivalently written as,

$$\rho(\pi) = \mathbb{E}_{(s,a,r)\sim d^\mathcal{D}}\left[w_{\pi/\mathcal{D}}(s,a)\cdot r\right], \tag{5}$$

where the correction terms are given by $w_{\pi/\mathcal{D}}(s,a) := \frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)}$, and our algorithm will focus on estimating these correction terms. Rather than relying on the assumption that $\mathcal{D}$ is the result of a single, known behavior policy, we instead make the following regularity assumption:

Assumption 1 (Reference distribution property)**.** *For any (s,a), $d^\pi(s,a) > 0$ implies $d^\mathcal{D}(s,a) > 0$. Furthermore, the correction terms are bounded by some finite constant C: $\left\|w_{\pi/\mathcal{D}}\right\|_\infty \le C$.*

[2] The Markovian requirement is necessary for TD methods. However, notably, IS methods do not depend on this assumption.

3 DualDICE

We now develop our algorithm, DualDICE, for estimating the discounted stationary distribution corrections $w_{\pi/\mathcal{D}}(s,a) = \frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)}$. In the OPE setting, one does not have explicit knowledge of the distribution $d^\mathcal{D}$, but rather only access to samples $\mathcal{D} = \{(s,a,r,s')\} \sim d^\mathcal{D}$. Similar to the TD methods described above, we also assume access to samples from the initial state distribution β. We begin by introducing a key result, which we will later derive and use as the crux for our algorithm.

3.1 The Key Idea

Consider optimizing a (bounded) function $\nu : S \times A \to \mathbb{R}$ for the following objective:

$$\min_{\nu: S\times A\to\mathbb{R}} J(\nu) := \frac{1}{2}\mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[\left(\nu - \mathcal{B}^\pi\nu\right)(s,a)^2\right] - (1-\gamma)\,\mathbb{E}_{s_0\sim\beta, a_0\sim\pi(s_0)}\left[\nu(s_0,a_0)\right], \quad (6)$$

where we use $\mathcal{B}^\pi$ to denote the expected Bellman operator with respect to policy π and zero reward: $\mathcal{B}^\pi\nu(s,a) = \gamma\mathbb{E}_{s'\sim T(s,a),a'\sim\pi(s')}[\nu(s',a')]$. The first term in equation 6 is the expected squared Bellman error with zero reward. This term alone would lead to a trivial solution $\nu^* \equiv 0$, which can be avoided by the second term that encourages $\nu^* > 0$. Together, these two terms result in an optimal ν^* that places some non-zero amount of Bellman residual at state-action pairs sampled from $d^\mathcal{D}$.

Perhaps surprisingly, as we will show, the Bellman residuals of ν^* are exactly the desired distribution corrections:

$$\left(\nu^* - \mathcal{B}^\pi\nu^*\right)(s,a) = w_{\pi/\mathcal{D}}(s,a). \quad (7)$$

This key result provides the foundation for our algorithm, since it provides us with a simple objective (relying only on samples from $d^\mathcal{D}$, β, π) which we may optimize in order to obtain estimates of the distribution corrections. In the text below, we will show how we arrive at this result. We provide one additional step which allows us to efficiently learn a parameterized ν with respect to equation 6. We then generalize our results to a family of similar algorithms and lastly present theoretical guarantees.

3.2 Derivation

A Technical Observation We begin our derivation of the algorithm for estimating $w_{\pi/\mathcal{D}}$ by presenting the following simple technical observation: For arbitrary scalars $m \in \mathbb{R}_{>0}, n \in \mathbb{R}_{\geq 0}$, the optimizer of the convex problem $\min_x J(x) := \frac{1}{2}mx^2 - nx$ is unique and given by $x^* = \frac{n}{m}$. Using this observation, and letting $\mathcal{C}$ be some bounded subset of $\mathbb{R}$ which contains $[0, C]$, one immediately sees that the optimizer of the following problem,

$$\min_{x: S\times A\to\mathcal{C}} J_1(x) := \frac{1}{2}\mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[x(s,a)^2\right] - \mathbb{E}_{(s,a)\sim d^\pi}\left[x(s,a)\right], \quad (8)$$

is given by $x^*(s,a) = w_{\pi/\mathcal{D}}(s,a)$ for any $(s,a) \in S \times A$. This result provides us with an objective that shares the same basic form as equation 6. The main distinction is that the second term relies on an expectation over d^π, which we do not have access to.

Change of Variables In order to transform the second expectation in equation 8 to be over the initial state distribution β, we perform the following change of variables: Let $\nu : S \times A \to \mathbb{R}$ be an arbitrary state-action value function that satisfies,

$$\nu(s,a) := x(s,a) + \gamma\mathbb{E}_{s'\sim T(s,a),a'\sim\pi(s')}[\nu(s',a')],\ \forall (s,a) \in S \times A. \quad (9)$$

Since $x(s,a) \in \mathcal{C}$ is bounded and $\gamma < 1$, the variable $\nu(s,a)$ is well-defined and bounded. By applying this change of variables, the objective function in 8 can be re-written in terms of ν, and this yields our previously presented objective from equation 6. Indeed, define,

$$\beta_t(s) := \Pr\left(s = s_t \mid s_0 \sim \beta, a_k \sim \pi(s_k), s_{k+1} \sim T(s_k,a_k) \text{ for } 0 \leq k < t\right),$$

to be the state visitation probability at step t when following π. Clearly, $\beta_0 = \beta$. Then,

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^\pi}\left[x(s,a)\right] &= \mathbb{E}_{(s,a)\sim d^\pi}\left[\nu(s,a) - \gamma\mathbb{E}_{s'\sim T(s,a),a'\sim\pi(s')}[\nu(s',a')]\right] \\
&= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{s\sim\beta_t,a\sim\pi(s)}\left[\nu(s,a) - \gamma\mathbb{E}_{s'\sim T(s,a),a'\sim\pi(s')}[\nu(s',a')]\right] \\
&= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{s\sim\beta_t,a\sim\pi(s)}\left[\nu(s,a)\right] - (1-\gamma)\sum_{t=0}^{\infty}\gamma^{t+1}\mathbb{E}_{s\sim\beta_{t+1},a\sim\pi(s)}\left[\nu(s,a)\right] \\
&= (1-\gamma)\mathbb{E}_{s\sim\beta,a\sim\pi(s)}\left[\nu(s,a)\right].
\end{aligned}$$

The Bellman residuals of the optimum of this objective give the desired off-policy corrections:

$$(\nu^* - \mathcal{B}^\pi \nu^*)(s,a) = x^*(s,a) = w_{\pi/\mathcal{D}}(s,a). \tag{10}$$

Equation 6 provides a promising approach for estimating the stationary distribution corrections, since the first expectation is over state-action pairs sampled from $d^\mathcal{D}$, while the second expectation is over β and actions sampled from π, both of which we have access to. Therefore, in principle we may solve this problem with respect to a parameterized value function ν, and then use the optimized ν^* to deduce the corrections. In practice, however, the objective in its current form presents two difficulties:

- The quantity $(\nu - \mathcal{B}^\pi \nu)(s,a)^2$ involves a conditional expectation inside a square. In general, when environment dynamics are stochastic and the action space may be large or continuous, this quantity may not be readily optimized using standard stochastic techniques. (For example, when the environment is stochastic, its Monte-Carlo sample gradient is generally biased.)
- Even if one has computed the optimal value ν^*, the corrections $(\nu^* - \mathcal{B}^\pi \nu^*)(s,a)$, due to the same argument as above, may not be easily computed, especially when the environment is stochastic or the action space continuous.

Exploiting Fenchel Duality We solve both difficulties listed above in one step by exploiting Fenchel duality [35]: Any convex function $f(x)$ may be written as $f(x) = \max_\zeta x \cdot \zeta - f^*(\zeta)$, where f^* is the Fenchel conjugate of f. In the case of $f(x) = \frac{1}{2}x^2$, the Fenchel conjugate is given by $f^*(\zeta) = \frac{1}{2}\zeta^2$. Thus, we may express our objective as,

$$\min_{\nu: S\times A\to\mathbb{R}} J(\nu) := \mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[\max_\zeta \ (\nu - \mathcal{B}^\pi\nu)(s,a)\cdot\zeta - \frac{1}{2}\zeta^2\right] - (1-\gamma)\,\mathbb{E}_{s_0\sim\beta, a_0\sim\pi(s_0)}\left[\nu(s_0,a_0)\right].$$

By the interchangeability principle [6, 34, 36], we may replace the inner max over scalar ζ to a max over functions $\zeta : S \times A \to \mathbb{R}$ and obtain a min-max saddle-point optimization:

$$\min_{\nu: S\times A\to\mathbb{R}} \max_{\zeta: S\times A\to\mathbb{R}} J(\nu,\zeta) := \mathbb{E}_{(s,a,s')\sim d^\mathcal{D}, a'\sim\pi(s')}\left[(\nu(s,a) - \gamma\nu(s',a'))\zeta(s,a) - \zeta(s,a)^2/2\right] - (1-\gamma)\,\mathbb{E}_{s_0\sim\beta, a_0\sim\pi(s_0)}\left[\nu(s_0,a_0)\right]. \tag{11}$$

Using the KKT condition of the inner optimization problem (which is convex and quadratic in ζ), for any ν the optimal value ζ^*_ν is equal to the Bellman residual, $\nu - \mathcal{B}^\pi\nu$. Therefore, the desired stationary distribution correction can then be found from the saddle-point solution (ν^*, ζ^*) of the minimax problem in equation 11 as follows:

$$\zeta^*(s,a) = (\nu^* - \mathcal{B}^\pi\nu^*)(s,a) = w_{\pi/\mathcal{D}}(s,a). \tag{12}$$

Now we finally have an objective which is well-suited for practical computation. First, unbiased estimates of both the objective and its gradients are easy to compute using stochastic samples from $d^\mathcal{D}$, β, and π, all of which we have access to. Secondly, notice that the min-max objective function in equation 11 is linear in ν and concave in ζ. Therefore in certain settings, one can provide guarantees on the convergence of optimization algorithms applied to this objective (see Section 3.4). Thirdly, the optimizer of the objective in equation 11 immediately gives us the desired stationary distribution corrections through the values of $\zeta^*(s,a)$, with no additional computation.

3.3 Extension to General Convex Functions

Besides a quadratic penalty function, one may extend the above derivations to a more general class of convex penalty functions. Consider a generic convex penalty function $f : \mathbb{R} \to \mathbb{R}$. Recall that $\mathcal{C}$ is a bounded subset of $\mathbb{R}$ which contains the interval $[0, C]$ of stationary distribution corrections. If $\mathcal{C}$ is contained in the range of f', then the optimizer of the convex problem, $\min_x J(x) := m \cdot f(x) - n$ for $\frac{n}{m} \in \mathcal{C}$, satisfies the following KKT condition: $f'(x^*) = \frac{n}{m}$. Analogously, the optimizer x^* of,

$$\min_{x: S\times A\to\mathcal{C}} J_1(x) := \mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[f(x(s,a))\right] - \mathbb{E}_{(s,a)\sim d^\pi}\left[x(s,a)\right], \tag{13}$$

satisfies the equality $f'(x^*(s,a)) = w_{\pi/\mathcal{D}}(s,a)$.

With change of variables $\nu := x + \mathcal{B}^\pi\nu$, the above problem becomes,

$$\min_{\nu: S\times A\to\mathbb{R}} J(\nu) := \mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[f((\nu - \mathcal{B}^\pi\nu)(s,a))\right] - (1-\gamma)\,\mathbb{E}_{s_0\sim\beta, a_0\sim\pi(s_0)}\left[\nu(s_0,a_0)\right]. \tag{14}$$

Applying Fenchel duality to f in this objective further leads to the following saddle-point problem:

$$\min_{\nu: S\times A\to\mathbb{R}} \max_{\zeta: S\times A\to\mathbb{R}} J(\nu,\zeta) := \mathbb{E}_{(s,a,s')\sim d^\mathcal{D}, a'\sim\pi(s')}\left[(\nu(s,a) - \gamma\nu(s',a'))\zeta(s,a) - f^*(\zeta(s,a))\right] - (1-\gamma)\,\mathbb{E}_{s_0\sim\beta, a_0\sim\pi(s_0)}\left[\nu(s_0,a_0)\right]. \tag{15}$$

By the KKT condition of the inner optimization problem, for any ν the optimizer ζ^*_ν satisfies,

$$f^{*\prime}(\zeta^*_\nu(s,a)) = (\nu - \mathcal{B}^\pi \nu)(s,a). \tag{16}$$

Therefore, using the fact that the derivative of a convex function f' is the inverse function of the derivative of its Fenchel conjugate $f^{*\prime}$, our desired stationary distribution corrections are found by computing the saddle-point (ζ^*, ν^*) of the above problem:

$$\zeta^*(s,a) = f'((\nu^* - \mathcal{B}^\pi \nu^*)(s,a)) = f'(x^*(s,a)) = w_{\pi/\mathcal{D}}(s,a). \tag{17}$$

Amazingly, despite the generalization beyond the quadratic penalty function $f(x) = \frac{1}{2}x^2$, the optimization problem in equation 15 retains all the computational benefits that make this method very practical for learning $w_{\pi/\mathcal{D}}(s,a)$: All quantities and their gradients may be unbiasedly estimated from stochastic samples; the objective is linear in ν and concave in ζ, thus is well-behaved; and the optimizer of this problem immediately provides the desired stationary distribution corrections through the values of $\zeta^*(s,a)$, without any additional computation.

This generalized derivation also provides insight into the initial technical result: It is now clear that the objective in equation 13 is the negative Fenchel dual (variational) form of the Ali-Silvey or f-divergence, which has been used in previous work to estimate divergence and data likelihood ratios [27]. In the case of $f(x) = \frac{1}{2}x^2$ (equation 8), this corresponds to a variant of the Pearson χ^2 divergence. Despite the similar formulations of our work and previous works using the same divergences to estimate data likelihood ratios [27], we emphasize that the aforementioned dual form of the f-divergence is not immediately applicable to estimation of off-policy corrections in the context of RL, due to the fact that samples from distribution d^π are unobserved. Indeed, our derivations hinged on two additional key steps: (1) the change of variables from x to $\nu := x + \mathcal{B}^\pi \nu$; and (2) the second application of duality to introduce ζ. Due to these repeated applications of duality in our derivations, we term our method *Dual stationary DIstribution Correction Estimation* (*DualDICE*).

3.4 Theoretical Guarantees

In this section, we consider the theoretical properties of DualDICE in the setting where we have a dataset formed by empirical samples $\{s_i, a_i, r_i, s'_i\}_{i=1}^N \sim d^\mathcal{D}$, $\{s^i_0\}_{i=1}^N \sim \beta$, and target actions $a'_i \sim \pi(s'_i), a^i_0 \sim \pi(s^i_0)$ for $i = 1, \ldots, N$.[3] We will use the shorthand notation $\hat{\mathbb{E}}_{d^\mathcal{D}}$ to denote an average over these empirical samples. Although the proposed estimator can adopt general f, for simplicity of exposition we restrict to $f(x) = \frac{1}{2}x^2$. We consider using an algorithm OPT (*e.g.*, stochastic gradient descent/ascent) to find optimal ν, ζ of equation 15 within some parameterization families $\mathcal{F}, \mathcal{H}$, respectively. We denote by $\hat{\nu}, \hat{\zeta}$ the outputs of OPT. We have the following guarantee on the quality of $\hat{\nu}, \hat{\zeta}$ with respect to the off-policy policy estimation (OPE) problem.

Theorem 2. *(Informal) Under some mild assumptions, the mean squared error (MSE) associated with using $\hat{\nu}, \hat{\zeta}$ for OPE can be bounded as,*

$$\mathbb{E}\left[\left(\hat{\mathbb{E}}_{d^\mathcal{D}}\left[\hat{\zeta}(s,a)\cdot r\right] - \rho(\pi)\right)^2\right] = \widetilde{\mathcal{O}}\left(\epsilon_{approx}(\mathcal{F},\mathcal{H}) + \epsilon_{opt} + \tfrac{1}{\sqrt{N}}\right), \tag{18}$$

where the outer expectation is with respect to the randomness of the empirical samples and OPT, ϵ_{opt} denotes the optimization error, and $\epsilon_{approx}(\mathcal{F},\mathcal{H})$ denotes the approximation error due to $\mathcal{F}, \mathcal{H}$.

The sources of estimation error are explicit in Theorem 2. As the number of samples N increases, the statistical error $N^{-1/2}$ approaches zero. Meanwhile, there is an implicit trade-off in $\epsilon_{approx}(\mathcal{F},\mathcal{H})$ and ϵ_{opt}. With flexible function spaces $\mathcal{F}$ and $\mathcal{H}$ (such as the space of neural networks), the approximation error can be further decreased; however, optimization will be complicated and it is difficult to characterize ϵ_{opt}. On the other hand, with linear parameterization of (ν, ζ), under some mild conditions, after T iterations we achieve provably fast rate, $\mathcal{O}(\exp(-T))$ for $OPT =$ SVRG and $\mathcal{O}\left(\frac{1}{T}\right)$ for $OPT =$ SGD, at the cost of potentially increased approximation error. See the Appendix for the precise theoretical results, proofs, and further discussions.

4 Related Work

Density Ratio Estimation Density ratio estimation is an important tool for many machine learning and statistics problems. Other than the naive approach, (i.e., the density ratio is calculated via estimating the densities in the numerator and denominator separately, which may magnify the estimation

[3] For the sake of simplicity, we consider the batch learning setting with *i.i.d.* samples as in [41]. The results can be easily generalized to single sample path with dependent samples (see Appendix).

error), various direct ratio estimators have been proposed [37], including the moment matching approach [13], probabilistic classification approach [3, 5, 33], and ratio matching approach [17, 27, 38]

The proposed DualDICE algorithm, as a direct approach for density ratio estimation, bears some similarities to ratio matching [27], which is also derived by exploiting the Fenchel dual representation of the f-divergences. However, compared to the existing direct estimators, the major difference lies in the requirement of the samples from the stationary distribution. Specifically, the existing estimators require access to samples from both $d^{\mathcal{D}}$ and d^{π}, which is impractical in the off-policy learning setting. Therefore, DualDICE is uniquely applicable to the more difficult RL setting.

Off-policy Policy Evaluation The problem of off-policy policy evaluation has been heavily studied in contextual bandits [8, 42, 45] and in the more general RL setting [10, 16, 20, 23, 28, 29, 30, 43, 44]. Several representative approaches can be identified in the literature. The Direct Method (DM) learns a model of the system and then uses it to estimate the performance of the evaluation policy. This approach often has low variance but its bias depends on how well the selected function class can express the environment dynamics. Importance sampling (IS) [31] uses importance weights to correct the mismatch between the distributions of the system trajectory induced by the target and behavior policies. Its variance can be unbounded when there is a big difference between the distributions of the evaluation and behavior policies, and grows exponentially with the horizon of the RL problem. Doubly Robust (DR) is a combination of DM and IS, and can achieve the low variance of DM and no (or low) bias of IS. Other than DM, all the methods described above require knowledge of the policy density ratio, and thus the behavior policy. Our proposed algorithm avoids this necessity.

5 Experiments

We evaluate our method applied to off-policy policy evaluation (OPE). We focus on this setting because it is a direct application of stationary distribution correction estimation, without many additional tunable parameters, and it has been previously used as a test-bed for similar techniques [21]. In each experiment, we use a behavior policy μ to collect some number of trajectories, each for some number of steps. This data is used to estimate the stationary distribution corrections, which are then used to estimate the average step reward, with respect to a target policy π. We focus our comparisons here to a TD-based approach (based on [12]) and weighted step-wise IS (as described in [21]), which we and others have generally found to work best relative to common IS variants [24, 31]. See the Appendix for additional results and implementation details.

We begin in a controlled setting with an evaluation agnostic to optimization issues, where we find that, absent these issues, our method is competitive with TD-based approaches (Figure 1). However, as we move to more difficult settings with complex environment dynamics, the performance of TD methods degrades dramatically, while our method is still able to provide accurate estimates (Figure 2). Finally, we provide an analysis of the optimization behavior of our method on a simple control task across different choices of function f (Figure 3). Interestingly, although the choice of $f(x) = \frac{1}{2}x^2$ is most natural, we find the empirically best performing choice to be $f(x) = \frac{2}{3}|x|^{3/2}$. All results are summarized for 20 random seeds, with median plotted and error bars at 25^{th} and 75^{th} percentiles.[4]

5.1 Estimation Without Function Approximation

We begin with a tabular task, the Taxi domain [7]. In this task, we evaluate our method in a manner agnostic to optimization difficulties: The objective 6 is a quadratic equation in ν, and thus may be solved by matrix operations. The Bellman residuals (equation 7) may then be estimated via an empirical average of the transitions appearing in the off-policy data. In a similar manner, TD methods for estimating the correction terms may also be solved using matrix operations [21]. In this controlled setting, we find that, as expected, TD methods can perform well (Figure 1), and our method achieves competitive performance. As we will see in the following results, the good performance of TD methods quickly deteriorates as one moves to more complex settings, while our method is able to maintain good performance, even when using function approximation and stochastic optimization.

5.2 Control Tasks

We now move on to difficult control tasks: A discrete-control task Cartpole and a continuous-control task Reacher [4]. In these tasks, observations are continuous, and thus we use neural network function

[4]The choice of plotting percentiles is somewhat arbitrary. Plotting mean and standard errors yields similar plots.

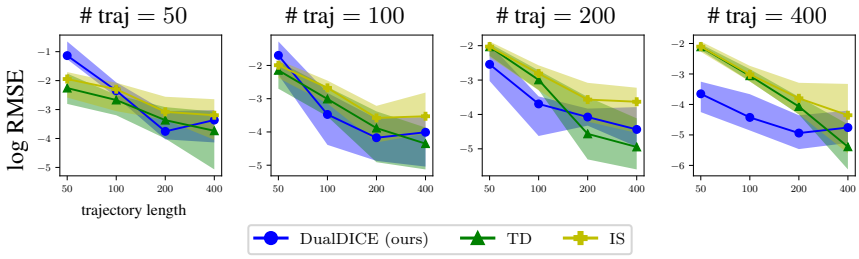

Figure 1: We perform OPE on the Taxi domain [7]. The plots show log RMSE of the estimator across different numbers of trajectories and different trajectory lengths (x-axis). For this domain, we avoid any potential issues in optimization by solving for the optimum of the objectives exactly using standard matrix operations. Thus, we are able to see that our method and the TD method are competitive with each other.

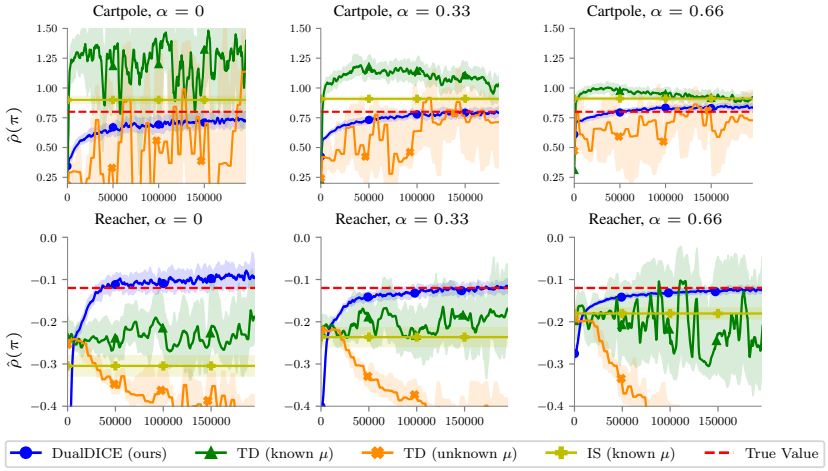

Figure 2: We perform OPE on control tasks. Each plot shows the estimated average step reward over training (x-axis is training step) and different behavior policies (higher α corresponds to a behavior policy closer to the target policy). We find that in all cases, our method is able to approximate these desired values well, with accuracy improving with a larger α. On the other hand, the TD method performs poorly, even more so when the behavior policy μ is unknown and must be estimated. While on Cartpole it can start to approach the desired value for large α, on the more complicated Reacher task (which involves continuous actions) its learning is too unstable to learn anything at all.

approximators with stochastic optimization. Figure 2 shows the results of our method compared to the TD method. We find that in this setting, DualDICE is able to provide good, stable performance, while the TD approach suffers from high variance, and this issue is exacerbated when we attempt to estimate μ rather than assume it as given. See the Appendix for additional baseline results.

5.3 Choice of Convex Function f

We analyze the choice of the convex function f. We consider a simple continuous grid task where an agent may move left, right, up, or down and is rewarded for reaching the bottom right corner of a square room. We plot the estimation errors of using DualDICE for off-policy policy evaluation on this task, comparing against different choices of convex functions of the form $f(x) = \frac{1}{p}|x|^p$. Interestingly, although the choice of $f(x) = \frac{1}{2}x^2$ is most natural, we find the empirically best performing choice to be $f(x) = \frac{2}{3}|x|^{3/2}$. Thus, this is the form of f we used in our experiments for Figure 2.

6 Conclusions

We have presented DualDICE, a method for estimating off-policy stationary distribution corrections. Compared to previous work, our method is agnostic to knowledge of the behavior policy used to collect the off-policy data and avoids the use of importance weights in its losses. These advantages have a profound empirical effect: our method provides significantly better estimates compared to TD methods, especially in settings which require function approximation and stochastic optimization.

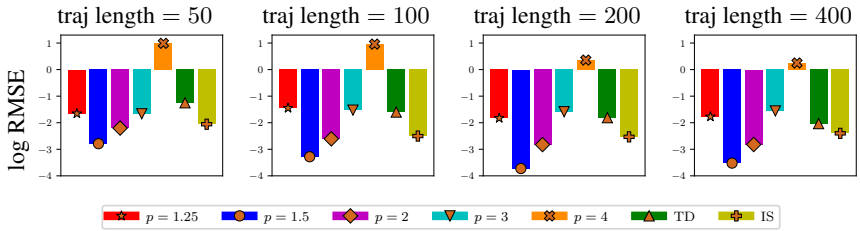

Figure 3: We compare the OPE error when using different forms of f to estimate stationary distribution ratios with function approximation, which are then applied to OPE on a simple continuous grid task. In this setting, optimization stability is crucial, and this heavily depends on the form of the convex function f. We plot the results of using $f(x) = \frac{1}{p}|x|^p$ for $p \in [1.25, 1.5, 2, 3, 4]$. We also show the results of TD and IS methods on this task for comparison. We find that $p = 1.5$ consistently performs the best, often providing significantly better results.

Future work includes (1) incorporating the DualDICE algorithm into off-policy training, (2) further understanding the effects of f on the performance of DualDICE (in terms of approximation error of the distribution corrections), and (3) evaluating DualDICE on real-world off-policy evaluation tasks.

Acknowledgments

We thank Marc Bellemare, Carles Gelada, and the rest of the Google Brain team for helpful thoughts and discussions.

References

[1] Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *arXiv preprint arXiv:1808.00177*, 2018.

[2] Richard Ernest Bellman. *Dynamic Programming*. Dover Publications, Inc., New York, NY, USA, 2003.

[3] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88. ACM, 2007.

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

[5] Kuang Fu Cheng, Chih-Kang Chu, et al. Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10(4):583–604, 2004.

[6] Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual embeddings. *arXiv preprint arXiv:1607.04579*, 2016.

[7] Thomas G Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, 13:227–303, 2000.

[8] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[9] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. *arXiv preprint arXiv:1802.03493*, 2018.

[10] Raphael Fonteneau, Susan A. Murphy, Louis Wehenkel, and Damien Ernst. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of Operations Research*, 208(1):383–416, 2013.

[11] Jianfeng Gao, Michel Galley, and Lihong Li. Neural approaches to Conversational AI. *Foundations and Trends in Information Retrieval*, 13(2–3):127–298, 2019.

[12] Carles Gelada and Marc G Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. *AAAI*, 2018.

[13] Arthur Gretton, Alex J Smola, Jiayuan Huang, Marcel Schmittfull, Karsten M Borgwardt, and Bernhard Schöllkopf. Covariate shift by kernel mean matching. In *Dataset shift in machine learning*, pages 131–160. MIT Press, 2009.

[14] Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1372–1383. JMLR. org, 2017.

[15] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[16] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 652–661, 2016.

[17] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul):1391–1445, 2009.

[18] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

[19] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306. ACM, 2011.

[20] Lihong Li, Rémi Munos, and Csaba Szepesvàri. Toward minimax off-policy value estimation. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 608–616, 2015.

[21] 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.

[22] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2019. To appear.

[23] A. Mahmood, H. van Hasselt, and R. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014.

[24] Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1077–1084. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[25] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[26] Susan A Murphy, Mark J van der Laan, James M Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

[27] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[28] C. Paduraru. *Off-policy Evaluation in Markov Decision Processes*. PhD thesis, McGill University, 2013.

[29] D. Precup, R. Sutton, and S. Dasgupta. Off-policy temporal difference learning with function approximation. In *Proceedings of the 18th International Conference on Machine Learning*, pages 417–424, 2001.

[30] D. Precup, R. Sutton, and S. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pages 759–766, 2000.

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

[32] Martin L Puterman. Markov decision processes: Discrete stochastic dynamic programming. 1994.

[33] Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

[34] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[35] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

[36] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[37] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[38] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

[39] Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*, volume 135.

[40] Richard S Sutton, A Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 17(1):2603–2631, 2016.

[41] Richard S Sutton, Csaba Szepesvári, Alborz Geramifard, and Michael Bowling. Dyna-style planning with linear function approximation and prioritized sweeping. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 528–536. AUAI Press, 2008.

[42] A. Swaminathan, A. Krishnamurthy, A. Agarwal, M. Dudík, J. Langford, D. Jose, and I. Zitouni. Off-policy evaluation for slate recommendation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3635–3645, 2017.

[43] P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2139–2148, 2016.

[44] P. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence off-policy evaluation. In *Proceedings of the 29th Conference on Artificial Intelligence*, 2015.

[45] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3589–3597. JMLR. org, 2017.