GradientDICE: Rethinking Generalized Offline Estimation of Stationary Values

Shangtong Zhang [1] **Bo Liu** [2] **Shimon Whiteson** [1]

Abstract

We present GradientDICE for estimating the density ratio between the state distribution of the target policy and the sampling distribution in off-policy reinforcement learning. GradientDICE fixes several problems of GenDICE (Zhang et al., 2020a), the state-of-the-art for estimating such density ratios. Namely, the optimization problem in GenDICE is *not* a convex-concave saddle-point problem once nonlinearity in optimization variable parameterization is introduced to ensure positivity, so any primal-dual algorithm is *not* guaranteed to converge or find the desired solution. However, such nonlinearity is essential to ensure the consistency of GenDICE even with a tabular representation. This is a fundamental contradiction, resulting from GenDICE's original formulation of the optimization problem. In GradientDICE, we optimize a different objective from GenDICE by using the Perron-Frobenius theorem and eliminating GenDICE's use of divergence, such that nonlinearity in parameterization is not necessary for GradientDICE, which is provably convergent under linear function approximation.

1. Introduction

A key challenge in reinforcement learning (RL, Sutton & Barto 2018) is off-policy evaluation (Precup et al., 2001; Maei, 2011; Jiang & Li, 2015; Sutton & Barto, 2018; Liu et al., 2018; Nachum et al., 2019; Zhang et al., 2020a), where we want to estimate the performance of a target policy (average reward in the continuing setting or expected total discounted reward in the episodic setting (Puterman, 2014)), from data generated by one or more behavior policies. Compared with on-policy evaluation (Sutton, 1988), which requires data generated by the target policy, off-policy evaluation is more flexible. We can evaluate a new policy with existing data in a replay buffer (Lin, 1992) without interacting with the environment again. We can also evaluate multiple target policies simultaneously when following a single behavior policy (Sutton et al., 2011).

One major challenge in off-policy evaluation is dealing with distribution mismatch: the state distribution of the target policy is different from the sampling distribution. This mismatch leads to divergence of the off-policy linear temporal difference learning algorithm (Baird, 1995; Tsitsiklis & Van Roy, 1997). Precup et al. (2001) address this issue with products of importance sampling ratios, which, however, suffer from a large variance. To correct for distribution mismatch without incurring a large variance, Hallak & Mannor (2017); Liu et al. (2018) propose to directly learn the density ratio between the state distribution of the target policy and the sampling distribution using function approximation.[3] Intuitively, this learned density ratio is a "marginalization" of the products of importance sampling ratios.

The density ratio learning algorithms from Hallak & Mannor (2017); Liu et al. (2018) require data generated by a *single known* behavior policy. Nachum et al. (2019) relax this constraint in DualDICE, which is compatible with *multiple unknown* behavior policies and *offline* training. DualDICE, however, copes well with only the total discounted reward criterion and cannot be used under the average reward criterion. In particular, DualDICE becomes unstable as the discounting factor grows towards 1 (Zhang et al., 2020a). To address the limitation of DualDICE, Zhang et al. (2020a) propose GenDICE, which is compatible with multiple unknown behavior policies and offline training *under both criteria*. Zhang et al. (2020a) show empirically that GenDICE achieves a new state-of-the-art in off-policy evaluation.

In this paper, we point out key problems with GenDICE. In particular, the optimization problem in GenDICE is *not* a convex-concave saddle-point problem (CCSP) once nonlinearity in optimization variable parameterization is introduced to ensure positivity, so any primal-dual algorithm is not guaranteed to converge or find the desired solution

[1]University of Oxford [2]Auburn University. Correspondence to: Shangtong Zhang <shangtong.zhang@cs.ox.ac.uk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

[3]Such density ratios are referred to as *stationary values* in Zhang et al. (2020a)

even with tabular representation. However, such positivity is essential to ensure the consistency of GenDICE. This is a fundamental contradiction, resulting from GenDICE's original formulation of the optimization problem.

Furthermore, we propose GradientDICE, which overcomes these problems. GradientDICE optimizes a different objective from GenDICE by using the Perron-Frobenius theorem (Horn & Johnson, 2012) and eliminating GenDICE's use of divergence. Consequently, nonlinearity in parameterization is not necessary for GradientDICE, which is provably convergent under linear function approximation. Finally, we provide empirical results demonstrating the advantages of GradientDICE over GenDICE and DualDICE.

2. Background

We use vectors and functions interchangeably when this does not confuse. For example, let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a function; we also use f to denote the corresponding vector in $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$. All vectors are column vectors.

We consider an infinite-horizon MDP with a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a transition kernel $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a discount factor $\gamma \in [0, 1)$, and an initial state distribution $\tilde{\mu}_0$. The initial state S_0 is sampled from $\tilde{\mu}_0$. At time step t, an agent at S_t selects an action A_t according to $\pi(\cdot|S_t)$, where $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the policy being followed by the agent. The agent then proceeds to the next state S_{t+1} according to $p(\cdot|S_t, A_t)$ and gets a reward R_{t+1} satisfying $\mathbb{E}[R_{t+1}] = r(S_t, A_t)$.

Similar to Zhang et al. (2020a), we consider two performance metrics for the policy π: the total discounted reward $\rho_\gamma(\pi) \doteq (1-\gamma)\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R_{t+1}]$ and the average reward $\rho(\pi) \doteq \lim_{T\to\infty} \frac{1}{T}\mathbb{E}[\sum_{t=1}^{T} R_t]$. When the Markov chain induced by π is ergodic, $\rho(\pi)$ is always well defined (Puterman, 2014). Throughout this paper, we implicitly assume this ergodicity whenever we consider $\rho(\pi)$. When considering $\rho_\gamma(\pi)$, we are interested in the normalized discounted state-action occupation $d_\gamma(s, a)$. Let $d_t^\pi(s, a) \doteq \Pr(S_t = s, A_t = a|\tilde{\mu}_0, p, \pi)$ be the probability of occupying the state-action pair (s, a) at the time step t following π. Then, we have $d_\gamma(s, a) \doteq (1-\gamma)\sum_{t=0}^{\infty} \gamma^t d_t^\pi(s, a)$. When considering $\rho(\pi)$, we are interested in the stationary state-action distribution $d_\gamma(s, a) \doteq \lim_{t\to\infty} d_t^\pi(s, a)$. To simplify notation, we extend the definition of $\rho_\gamma(\pi)$ and $d_\gamma(s, a)$ from $\gamma \in [0, 1)$ to $\gamma \in [0, 1]$ by defining $\rho_1(\pi) \doteq \rho(\pi)$ and $d_1(s, a) \doteq d(s, a)$. It follows that for any $\gamma \in [0, 1]$, we have $\rho_\gamma(\pi) = \mathbb{E}_{(s,a)\sim d_\gamma}[r(s, a)]$.

We are interested in estimating $\rho_\gamma(\pi)$ without executing the policy π. Similar to Zhang et al. (2020a), we assume access to a fixed dataset $\mathcal{D} \doteq \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$. Here the state-action pair (s_i, a_i) is sampled from an unknown distribution $d_\mu : \mathcal{S} \times \mathcal{A} \to [0, 1]$, which may result from multiple unknown behavior policies. The reward r_i satisfies $\mathbb{E}[r_i] = r(s_i, a_i)$. The successor state s'_i is sampled from $p(\cdot|s_i, a_i)$. As $\rho_\gamma(\pi) = \mathbb{E}_{(s,a)\sim d_\mu}[\frac{d_\gamma(s,a)}{d_\mu(s,a)} r(s, a)]$, one possible approach for estimating $\rho_\gamma(\pi)$ is to learn the density ratio $\tau_*(s, a) \doteq \frac{d_\gamma(s,a)}{d_\mu(s,a)}$ directly.

We assume $d_\mu(s, a) > 0$ for every (s, a) and use $D \in \mathbb{R}^{N_{sa}\times N_{sa}}(N_{sa} \doteq |\mathcal{S}| \times |\mathcal{A}|)$ to denote a diagonal matrix whose diagonal entry is d_μ. Let $\mu_0(s, a) \doteq \tilde{\mu}_0(s)\pi(a|s)$, Zhang et al. (2020a) show

$$D\tau_* = \mathcal{T}\tau_*,$$

where the operator $\mathcal{T}$ is defined as

$$\mathcal{T}y \doteq (1-\gamma)\mu_0 + \gamma P_\pi^\top Dy,$$

and $P_\pi \in \mathbb{R}^{N_{sa}\times N_{sa}}$ is the state-action pair transition matrix, i.e., $P_\pi((s, a), (s', a')) \doteq p(s'|s, a)\pi(a'|s')$. The operator $\mathcal{T}$ is similar to the Bellman operator but in the reverse direction. Similar ideas have been explored by Wang et al. (2007; 2008); Hallak & Mannor (2017); Liu et al. (2018); Gelada & Bellemare (2019). As $D\tau_*$ is a probability measure, Zhang et al. (2020a) propose to compute τ_* by solving the following optimization problem:

$$\min_{\tau\in\mathbb{R}^{N_{sa}}, \tau\succeq 0} D_\phi(\mathcal{T}\tau||D\tau) \quad \text{s.t. } d_\mu^\top \tau = 1, \tag{1}$$

where $\tau \succeq 0$ is elementwise greater or equal, D_ϕ is an f-divergence (Nowozin et al., 2016) associated with a convex, lower semi-continuous generator function $\phi : \mathbb{R}_+ \to \mathbb{R}$ with $\phi(1) = 0$. Let q_1, q_2 be two probability measures; we have $D_\phi(q_1||q_2) \doteq \sum_x q_2(x)\phi(\frac{q_1(x)}{q_2(x)})$. The f-divergence is used mainly for the ease of optimization but see Zhang et al. (2020a) for discussion of other possible divergences. Due to the difficulty in solving the constrained problem (1) directly, Zhang et al. (2020a) propose to solve the following problem instead:

$$\min_{\tau\in\mathbb{R}^{N_{sa}}, \tau\succeq 0} D_\phi(\mathcal{T}\tau||D\tau) + \tfrac{\lambda}{2}(d_\mu^\top \tau - 1)^2, \tag{2}$$

where $\lambda > 0$ is a constant. We have

Lemma 1. *(Zhang et al., 2020a) For any constant $\lambda > 0$, τ is optimal for the problem (2) iff $\tau = \tau_*$.*

To make the optimization tractable and address a double sampling issue, Zhang et al. (2020a) rewrite $\phi(x)$ as $\phi(x) = \max_{f\in\mathbb{R}} xf - \frac{1}{2}\phi^*(f)$, where ϕ^* is the Fenchel conjugate of ϕ, and use the interchangeability principle for interchanging maximization and expectation (Shapiro et al., 2014; Dai et al., 2016), yielding the following problem, which is equivalent to (2):

$$\min_{\tau\in\mathbb{R}^{N_{sa}}, \tau\succeq 0} \max_{f\in\mathbb{R}^{N_{sa}}, \eta\in\mathbb{R}} J(\tau, f, \eta), \tag{3}$$

where

$$J(\tau, f, \eta)$$
$$\doteq (1-\gamma)\mathbb{E}_{\mu_0}[f(s,a)] + \gamma \mathbb{E}_p[\tau(s,a) f(s',a')] -$$
$$\mathbb{E}_{d_\mu}[\tau(s,a)\phi^*(f(s,a))] + \lambda\Big(\mathbb{E}_{d_\mu}[\eta\tau(s,a) - \eta] - \frac{\eta^2}{2}\Big).$$

Here $\mathbb{E}_{\mu_0}, \mathbb{E}_{d_\mu}, \mathbb{E}_p$ are shorthand for $\mathbb{E}_{(s,a)\sim\mu_0}$, $\mathbb{E}_{(s,a)\sim d_\mu}$, $\mathbb{E}_{(s,a)\sim d_\mu, s'\sim p(\cdot|s,a), a'\sim\pi(\cdot|s')}$ respectively. Zhang et al. (2020a) show J is convex in τ and concave in η, f, i.e., (3) is a convex-concave saddle-point problem. Zhang et al. (2020a) therefore use a primal-dual algorithm (i.e, perform stochastic gradient ascent on η, f and stochastic gradient descent on τ) to find the saddle-point, yielding GENeralized stationary DIstribution Correction Estimation (GenDICE).

3. Problems with GenDICE

3.1. Use of Divergences as the Objective

Zhang et al. (2020a) proposes to consider a family of divergences, the f-divergences. However, f-divergences are defined between probability measures. So D_ϕ in (2) implicitly requires its arguments to be valid probability measures. Consequently, (2) still has the implicit constraint that $d_\mu^\top \tau = 1$. The main motivation for Zhang et al. (2020a) to transform (1) into (2) is to get rid of this equality constraint. By using divergences, they do not really get rid of it. When this implicit constraint is considered, the problem (2) is still hard to optimize, as discussed in Zhang et al. (2020a).

We can, of course, just ignore this implicit constraint and interpret D_ϕ as a generic function instead of a divergence. Namely, we do not require its arguments to be valid probability measures. In this scenario, however, there is no guarantee that D_ϕ is always nonnegative, which plays a central role in proving Lemma 1's claim that GenDICE is consistent. Consider, for example, the KL-divergence, where $\phi(x) = x\log(x)$. If $q_2(x) > q_1(x) > 0$ holds for all x (which is impossible when q_1 and q_2 are probability measures), clearly we have $D_\phi(q_1||q_2) < 0$. While Zhang et al. (2020a) propose not to use KL divergence due to numerical instability, here we provide a more principled explanation that if KL divergence is used, Lemma 1 does not necessarily hold. Zhang et al. (2020a) propose to use χ^2-divergence instead. Fortunately, χ^2-divergence has the property that $q_1(x) > 0 \wedge q_2(x) > 0$ implies $D_\phi(q_1||q_2) \geq 0$, even if q_1 and q_2 are not probability measures. This property ensures Lemma 1 holds even we just consider D_ϕ as a generic function instead of a divergence. But not all divergences have this property. Moreover, even if χ^2-divergence is considered, $\tau \succeq 0$ is still necessary for Lemma 1 to hold. This requirement ($\tau \succeq 0$) is also problematic, as discussed in the next section.

To summarize, we argue that *f-divergence is not a good choice to form the optimization objective for density ratio learning*.

3.2. Use of Primal-Dual Algorithms as the Solver

We assume τ, f are parameterized by $\theta^{(1)}, \theta^{(2)}$. As (3) requires $\tau(s,a) \geq 0$, Zhang et al. (2020a) propose to add extra nonlinearity, e.g., $(\cdot)^2, \log(1+\exp(\cdot))$, or $\exp(\cdot)$, in the parameterization of τ. Plugging the approximation in (3) yields $\min_{\theta^{(1)}} \max_{\theta^{(2)},\eta} J(\tau_{\theta^{(1)}}, \eta, f_{\theta^{(2)}})$. Here $\tau_{\theta^{(1)}}$ and $f_{\theta^{(2)}}$ emphasize that τ and f are parameterized functions.

There is now a contradiction. On the one hand, $J(\tau_{\theta^{(1)}}, \eta, f_{\theta^{(2)}})$ is *not* necessarily CCSP when nonlinearity is introduced in the parameterization. In the definition of J in (3), the sign of τ depends on f and η. Unless τ is linear in $\theta^{(1)}$, the convexity of J w.r.t. $\theta^{(1)}$ is in general hard to analyze (Boyd & Vandenberghe, 2004), even we just add $(\cdot)^2$ after a linear parameterization. Although Zhang et al. (2020a) demonstrate great empirical success from a primal-dual algorithm, this optimization procedure is *not* theoretically justified as $J(\tau_{\theta^{(1)}}, \eta, f_{\theta^{(2)}})$ is not necessarily a convex-concave function. On the other hand, if we do not apply any nonlinearity in $\theta^{(1)}$, there is no guarantee that $\tau_{\theta^{(1)}}(s,a) > 0$ even with a tabular representation. Then Lemma 1 does not necessarily hold, and GenDICE is not necessarily consistent.

To summarize, *applying the primitive primal-dual algorithm for the GenDICE objective is not theoretically justified, even with a tabular representation*.

3.3. Projection and Self-Normalization

Besides applying nonlinearity, one may also consider projection to account for the constraint $\tau_{\theta^{(1)}}(s,a) > 0$, i.e., we project $\theta^{(1)}$ back to $\Theta \doteq \{\theta^{(1)} | \tau_{\theta^{(1)}}(s,a) > 0\}$. One direct instantiation of this idea is Projected Stochastic Gradient Descent (PSGD). With nonlinear function approximation, it is not clear how to achieve this. With tabular or linear representation, such projection reduces to inequality-constrained quadratic programming, solving which usually involves subroutine numerical optimization, making it very computationally expensive. Moreover, the number of constraints grows linearly w.r.t. the number of states (not state features), indicating such projection does not scale well. Stochastic Mirror Descent (SMD, Beck & Teboulle 2003) with generalized KL-divergence is also a possible way to to achieve such a projection. SMD, as well as PSGD, walks $\theta^{(1)}$ in the positive orthant. To ensure $\tau_{\theta^{(1)}}(s,a) > 0$, they also require positive features. However, it is possible that the oracle τ lies in the feature space but the optimal $\theta^{(1)}$ does not lie in the positive orthant, indicating SMD and PSGD can yield arbitrarily large optimization errors.

Self-normalization (Liu et al., 2018; Zhang et al., 2020a) is also an approach to ensure nonlinearity, where we normalize the τ prediction over all possible state-action pairs. Recently, Mousavi et al. (2020) propose an objective with Maximum Mean Discrepancy and use self-normalization. However, self-normalization usually generates biased solutions and is computationally prohibitive for large datasets (see Appendix E.3 in Zhang et al. (2020a)). Moreover, self-normalization in general yields non-convex objectives even with tabular or linear representation (e.g., the objective in Mousavi et al. (2020) is non-convex), rendering difficulties in optimization.

To summarize, we argue that *neither projection nor self-normalization is a theoretically justified solution for the problems of GenDICE.*

3.4. A Hard Example for GenDICE

Figure 1. A single-state MDP.

We now provide a concrete example to demonstrate the defects of GenDICE. We consider a single state MDP (Figure 1) with two actions, both of which lead to that single state. We set $\gamma = 1, d_\mu(s_0, a_1) = d_\mu(s_0, a_2) = \pi(a_1|s_0) = \pi(a_2|s_0) = 0.5$. Therefore, we have $\mu_0(s_0, a_1) = \mu_0(s_0, a_2) = 0.5$. Under this setting, it is easy to verify that $\tau_* = [1, 1]^\top$. We now instantiate (3) with a χ^2-divergence and $\lambda = 1$ as recommended by Zhang et al. (2020a), where $\phi^*(x) = x + \frac{x^2}{4}$. To solve (3), we need $\tau \succeq 0$. As suggested by Zhang et al. (2020a), we use $(\cdot)^2$ nonlinearity. Namely, we define $\tau(s_0, a_1) = \tau_1^2, \tau(s_0, a_2) = \tau_2^2, f(s_0, a_1) = f_1, f(s_0, a_2) = f_2$. Now our optimization variables are $\tau_1, \tau_2, f_1, f_2, \eta$. It is easy to verify that at the point $(\tau_1, \tau_2, f_1, f_2, \eta) = (0, 0, 0, 0, -1)$, we have $\frac{\partial J(\tau,f,\eta)}{\partial \tau_1} = \frac{\partial J(\tau,f,\eta)}{\partial \tau_2} = \frac{\partial J(\tau,f,\eta)}{\partial f_1} = \frac{\partial J(\tau,f,\eta)}{\partial f_2} = \frac{\partial J(\tau,f,\eta)}{\partial \eta} = 0$, indicating GenDICE stops at this point if the true gradient is used. However, $[0, 0]^\top$ is obviously not the optimum. Details of the computation are provided in the appendix. This suboptimality results from the fact that once nonlinearity is introduced in $J(\tau, f, \eta)$, it is not convex-concave even with a tabular representation. Although this example is trivial to solve, it numerically verifies the problems of GenDICE.

4. GradientDICE

As discussed above, the problems with GenDICE come mainly from the formulation of (2), namely the constraint $\tau \succeq 0$ and the use of the divergence D_ϕ. We eliminate both by considering the following problem instead:

$$\min_{\tau \in \mathbb{R}^{N_{sa}}} L(\tau) \doteq \frac{1}{2}||\mathcal{T}\tau - D\tau||^2_{D^{-1}} + \frac{\lambda}{2}(d_\mu^\top \tau - 1)^2, \quad (4)$$

where $\lambda > 0$ is a constant and $||y||^2_\Xi$ stands for $y^\top \Xi y$. Readers familiar with Gradient TD methods (Sutton et al., 2009a;b) or residual gradients (Baird, 1995) may find the first term of this objective similar to the Mean Squared Bellman Error (MSBE). However, while in MSBE the norm is induced by D, we consider a norm induced by D^{-1}. This norm is carefully designed and provides expectations that we can sample from, which will be clear once $L(\tau)$ is expanded (see Eq (5) below). Remarkably, we have:

Theorem 1. *τ is optimal for (4) iff $\tau = \tau_*$.*

Proof. Sufficiency: Obviously τ_* is optimal.
Necessity: (a) $\gamma < 1$: In this case $I - \gamma P_\pi^\top$ is nonsingular. The linear system $D\tau - \mathcal{T}\tau = 0$ has only one solution, we must have $\tau = \tau_*$. (b) $\gamma = 1$: If τ is optimal, we have $d_\mu^\top \tau = 1$ and $D\tau = P_\pi^\top D\tau$, i.e., $D\tau$ is a left eigenvector of P_π associated with the Perron-Frobenius eigenvalue 1. Note d_γ is also a left eigenvector of P_π associated with the eigenvalue 1. According to the Perron-Frobenius theorem for nonnegative irreducible matrices (Horn & Johnson, 2012), the left eigenspace of the Perron-Frobenius eigenvalue is 1-dimensional. Consequently, there exists a scalar α such that $D\tau = \alpha d_\gamma$. On the other hand, $\alpha = \alpha 1^\top d_\gamma = 1^\top D\tau = d_\mu^\top \tau = 1$, implying $D\tau = d_\gamma$, i.e., $\tau = \tau_*$. □

Remark 1. *Unlike the problem formulation in Zhang et al. (2020a) (see (1) and (2)), we do not use $\tau \succeq 0$ as a constraint and can still guarantee there is no degenerate solution. Eliminating this constraint is key to eliminating nonlinearity. Although the Perron-Frobenius theorem can also be used in the formulation of Zhang et al. (2020a), their use of the divergence D_ϕ still requires $\tau \succeq 0$.*

With $\delta = \mathcal{T}\tau - D\tau$, we have

$$\begin{aligned} L(\tau) &\doteq \frac{1}{2}\mathbb{E}_{d_\mu}[\Big(\frac{\delta(s,a)}{d_\mu(s,a)}\Big)^2] + \frac{\lambda}{2}(d_\mu^\top \tau - 1)^2 \quad (5) \\ &= \max_{f \in \mathbb{R}^{N_{sa}}} \mathbb{E}_{d_\mu}[\frac{\delta(s,a)}{d_\mu(s,a)} f(s,a) - \frac{1}{2} f(s,a)^2] \\ &\quad + \lambda \max_{\eta \in \mathbb{R}} \big(\mathbb{E}_{d_\mu}[\eta\tau(s,a) - \eta] - \frac{\eta^2}{2}\big), \end{aligned}$$

where the equality comes also from the Fenchel conjugate and the interchangeability principle as in Zhang et al.

(2020a). We, therefore, consider the following problem

$$\begin{aligned}
&\min_{\tau\in\mathbb{R}^{N_{sa}}} \max_{f\in\mathbb{R}^{N_{sa}},\eta\in\mathbb{R}} \Big(L(\tau,\eta,f)\\
\doteq&\mathbb{E}_{d_\mu}[\tfrac{\delta(s,a)}{d_\mu(s,a)}f(s,a)-\tfrac{1}{2}f(s,a)^2]\\
&+\lambda\big(\eta(\mathbb{E}_{d_\mu}[\tau(s,a)-1])-\tfrac{\eta^2}{2}\big)\\
=&(1-\gamma)\mathbb{E}_{\mu_0}[f(s,a)]+\gamma\mathbb{E}_p[\tau(s,a)f(s',a')]\\
&-\mathbb{E}_{d_\mu}[\tau(s,a)f(s,a)]-\frac{1}{2}\mathbb{E}_{d_\mu}[f(s,a)^2]\\
&+\lambda\big(\mathbb{E}_{d_\mu}[\eta\tau(s,a)-\eta]-\tfrac{\eta^2}{2}\big)\Big).
\end{aligned}$$

Here the equality comes from the fact that

$$\mathbb{E}_p[\tau(s,a)f(s',a')]=\textstyle\sum_{s',a'}(P_\pi^\top D\tau)(s',a')f(s',a').$$

This problem is an *unconstrained* optimization problem and L is convex (linear) in τ and concave in f,η. Assuming τ, f is parameterized by w,κ respectively and including ridge regularization for w for reasons that will soon be clear, we consider the following problem

$$\min_w \max_{\eta,\kappa} L(\tau_w,\eta,f_\kappa)+\tfrac{\xi}{2}||w||^2, \tag{6}$$

where $\xi\geq 0$ is a constant. When a linear architecture is considered for τ_w and f_κ, the problem (6) is CCSP. Namely, it is convex in w and concave in κ,η. We use $X\in\mathbb{R}^{N_{sa}\times K}$ to denote the feature matrix, each row of which is $x(s,a)$, where $x:\mathcal{S}\times\mathcal{A}\to\mathbb{R}^K$ is the feature function. Assuming $\tau_w\doteq Xw, f_\kappa\doteq X\kappa$, we perform gradient descent on w and gradient ascent on η,κ. As we use techniques similar to Gradient TD methods to prove the convergence of our new algorithm, we term it Gradient stationary DIstribution Correction Estimation (GradientDICE):

$$\begin{aligned}
\delta_t&\leftarrow(1-\gamma)x_{0,t}+\gamma x_t^\top w_t x_t'-x_t^\top w_t x_t\\
\kappa_{t+1}&\leftarrow\kappa_t+\alpha_t(\delta_t-x_t^\top\kappa_t x_t)\\
\eta_{t+1}&\leftarrow\eta_t+\alpha_t\lambda(x_t^\top w_t-1-\eta_t)\\
w_{t+1}&\leftarrow w_t-\alpha_t\Big(\gamma x_t'^\top\kappa_t x_t-x_t^\top\kappa_t x_t+\lambda\eta_t x_t+\xi w_t\Big).
\end{aligned}$$

Here $x_{0,t}\doteq x(s_{0,t},a_{0,t}), x_t\doteq x(s_t,a_t), x_t'\doteq(s_t',a_t')$, $(s_{0,t},a_{0,t})\sim\mu_0, (s_t,a_t,s_t',a_t')\sim p$ (c.f. $\mathbb{E}_p$ in (3)), $\{\alpha_t\}$ is a sequence of deterministic nonnegative nonincreasing learning rates satisfying the Robin-Monro's condition (Robbins & Monro, 1951), i.e., $\sum_t\alpha_t=\infty, \sum_t\alpha_t^2<\infty$.

4.1. Convergence Analysis

Let $d_t^\top\doteq[\kappa_t^\top, w_t^\top,\eta_t]$, we rewrite the GradientDICE updates as $d_{t+1}\doteq d_t+\alpha_t(G_{t+1}d_t+g_{t+1})$, where

$$G_{t+1}\doteq\begin{bmatrix}-x_t^\top x_t & -(x_t-\gamma x_t')x_t^\top & 0\\ x_t(x_t^\top-\gamma x_t'^\top) & -\xi I & -\lambda x_t\\ 0 & \lambda x_t^\top & -\lambda\end{bmatrix},$$
$$g_{t+1}\doteq\begin{bmatrix}(1-\gamma)x_{0,t}\\0\\-\lambda\end{bmatrix}.$$

Defining $A\doteq X^\top(I-\gamma P_\pi^\top)DX, C\doteq X^\top DX$, the limiting behavior of GradientDICE is governed by

$$G\doteq\mathbb{E}_p[G_t]=\begin{bmatrix}-C & -A & 0\\ A^\top & -\xi I & -\lambda X^\top d_\mu\\ 0 & \lambda d_\mu^\top X & -\lambda\end{bmatrix},$$
$$g\doteq\mathbb{E}_{\mu_0}[g_t]=\begin{bmatrix}(1-\gamma)X^\top\mu_0\\0\\-\lambda\end{bmatrix}.$$

Assumption 1. *X has linearly independent columns.*

Assumption 2. *A is nonsingular or $\xi>0$.*

Assumption 3. *The features $\{x_{0,t},x_t,x_t'\}$ have uniformly bounded second moments.*

Remark 2. *Assumption 1 ensures C is strictly positive definite. When $\gamma<1$, it is common to assume A is nonsingular (Maei, 2011), the ridge regularization (i.e., $\xi>0$) is then optional. When $\gamma=1$, A can easily be singular (e.g., in a tabular setting). We, therefore, impose the extra ridge regularization. Assumption 3 is commonly used in Gradient TD methods (Maei, 2011). Assumptions (1 - 3) are also used in previous density ratio learning literature for analyzing the optimization error, either explicitly (e.g., the Assumption 8 in Appendix D.4 in Nachum et al. (2019)) or implicitly (e.g., Appendix C.3 in Zhang et al. (2020a)).*

Theorem 2. *Under Assumptions (1-3), we have*

$$\lim_{t\to\infty} d_t=-G^{-1}g \quad \textit{almost surely}.$$

We provide a detailed proof of Theorem 2 in the appendix, which is inspired by Sutton et al. (2009a). One key step in the proof is to show that the real parts of all eigenvalues of G are strictly negative. The G in Sutton et al. (2009a) satisfies this condition easily. However, for our G to satisfy this condition when $\gamma=1$, we must have $\xi>0$, which motivates the use of ridge regularization.

With simple block matrix inversion expanding G^{-1}, we

have $\lim_{t\to\infty} w_t = w_{\infty,\xi}$, where

$$\begin{aligned} w_{\infty,\xi} &\doteq (1-\gamma)\Xi A^\top C^{-1} X^\top \mu_0 \\ &\quad + \lambda\beta^{-1} z[1-(1-\gamma)z^\top A^\top C^{-1} X^\top \mu_0] \\ \Xi &\doteq (\xi I + A^\top C^{-1} A)^{-1}, \\ z &\doteq \Xi X^\top d_\mu, \quad \beta \doteq 1 + \lambda d_\mu^\top X \Xi X^\top d_\mu. \end{aligned}$$

The maximization step in (6) is quadratic (with linear function approximation) and thus can be solved analytically. Simple algebraic manipulation together with Assumption 1 shows that this quadratic problem has a unique optimizer for all $\gamma \in [0,1]$. Plugging the analytical solution for the maximization step in (6), the KKT conditions then state that the optimizer $w_{*,\xi}$ for the minimization step must satisfy $A_{*,\xi} w_{*,\xi} = b_*$, where

$$\begin{aligned} A_{*,\xi} &\doteq A^\top C^{-1} A + \lambda X^\top d_\mu d_\mu^\top X + \xi I, \\ b_* &\doteq (1-\gamma) A^\top C^{-1} X^\top \mu_0 + \lambda X^\top d_\mu. \end{aligned}$$

Assumption 2 ensures $A_{*,\xi}$ is nonsingular. Using the Sherman-Morrison formula (Sherman & Morrison, 1950), it is easy to verify $w_{*,\xi} = w_{\infty,\xi}$. For a quick sanity check, it is easy to verify that $w_{\infty,0} = \tau_*$ holds when $\gamma < 1$, $X = I$, and $\xi = 0$, using the fact $(1-\gamma)1^\top (I - \gamma P_\pi^\top)\mu_0 = 1$.

4.2. Consistency Analysis

To ensure convergence, we require ridge regularization in (6) for the setting $\gamma = 1$. The asymptotic solution $w_{\infty,\xi}$ is therefore biased. We now study the regularization path consistency for the setting $\gamma = 1$, i.e., we study the behavior of $w_{\infty,\xi}$ when ξ approaches 0.

Case 1: $\tau_* \in col(X)$. Here $col(\cdot)$ indicates the column space. As X has linearly independent columns (Assumption 1), we use w_* to denote the unique w satisfying $Xw = \tau_*$. As $\gamma = 1$, A can be singular. Hence both $w_{\infty,0}$ and $A_{*,0}^{-1}$ can be ill-defined. We now show under some regularization, we still have the desired consistency. As $A^\top C^{-1} A$ is always positive semidefinite, we consider its eigendecomposition $A^\top C^{-1} A = Q^\top \Lambda Q$, where Q is an orthogonal matrix, $\Lambda \doteq diag([\lambda_1, \cdots, \lambda_r, 0, \cdots, 0])$, r is the rank of $A^\top C^{-1} A$, $\lambda_i > 0$ are eigenvalues. Let $u \doteq Q X^\top d_\mu$, we have

Proposition 1. *Assuming $XC^{-1}X^\top$ is positive definite, $||u_{r+1:N_{sa}}|| \neq 0$, then $\lim_{\xi\to 0} w_{\infty,\xi} = w_*$, where $u_{i:j}$ denotes the vector consisting of the elements indexed by $i, i+1, \ldots, j$ in the vector u.*

Proof. According to the Perron-Frobenius theorem (c.f. the proof of Theorem 1), it suffices to show

$$\begin{aligned} &\lim_{\xi\to 0} L_1(w_{\infty,\xi}) = \lim_{\xi\to 0} L_2(w_{\infty,\xi}) = 0, \\ &L_1(w_{\infty,\xi}) \doteq d_\mu^\top X w_{\infty,\xi} - 1, \\ &L_2(w_{\infty,\xi}) \doteq ||DXw_{\infty,\xi} - P_\pi^\top X D w_{\infty,\xi}||^2_{XC^{-1}X^\top}, \end{aligned}$$

as w_* is the only w satisfying $L_1(w) = L_2(w) = 0$. With the eigendecomposition of $A^\top C^{-1} A$, we can compute Ξ explicitly. Simple algebraic manipulation then yields

$$\begin{aligned} L_1(w_{\infty,\xi}) &= \tfrac{\lambda u^\top \Lambda_\xi u}{1+\lambda u^\top \Lambda_\xi u} - 1, \\ L_2(w_{\infty,\xi}) &= \tfrac{\lambda^2 u^\top \Lambda_\xi u}{(1+\lambda u^\top \Lambda_\xi u)^2} + \tfrac{\lambda^2 \xi u^\top \Lambda_\xi^2 u}{(1+\lambda u^\top \Lambda_\xi u)^2}, \end{aligned}$$

where $\Lambda_\xi \doteq diag([\frac{1}{\xi+\lambda_1}, \ldots, \frac{1}{\xi+\lambda_r}, \frac{1}{\xi}, \cdots, \frac{1}{\xi}])$. The desired limits then follow from the L'Hopital's rule. □

Remark 3. *The assumption $||u_{r+1:N_{sa}}|| \neq 0$ is not restrictive as it is independent of learnable parameters and mainly controlled by features. Requiring $XC^{-1}X^\top$ to be positive definite is more restrictive, but it holds at least for the tabular setting (i.e., $X = I$). The difficulty of the setting $\gamma = 1$ comes mainly from the fact that the objective of the minimization step in the problem (6) is no longer strictly convex when $\xi = 0$ (i.e., $A_{*,0}$ can be singular). Thus there may be multiple optima for this minimization step, only one of which is w_*. Extra domain knowledge (e.g., assumptions in the proposition statement) is necessary to ensure the regularization path converges to the desired optimum. We provide a sufficient condition here and leave the analysis of necessary conditions for future work.*

Case 2: $\tau_* \notin col(X)$. In this scenario, it is not clear how to define w_*. The minimization step in (6) can have multiple optima, and it is not clear which one is the best. To analyze this scenario, we need to explicitly define projection in the optimization objective like Mean Squared Projected Bellman Error (Sutton et al., 2009a), instead of using an MSBE-like objective. We leave this for future work.

4.3. Finite Sample Analysis

We now provide a finite sample analysis for a variant of GradientDICE, *Projected GradientDICE* (Algorithm 1), where we introduce projection and iterates average, akin to Nemirovski et al. (2009); Liu et al. (2015). Intuitively, Projected GradientDICE groups the κ, η in GradientDICE into $y^\top = [\kappa^\top, \eta]$. Precisely, we have $y_t \in \mathbb{R}^{K+1}, w_t \in \mathbb{R}^K$,

$$\begin{aligned} G_{1,t} &\doteq \begin{bmatrix} -x_t^\top x_t & 0 \\ 0 & -\lambda \end{bmatrix}, G_{2,t} \doteq \begin{bmatrix} -(x_t - \gamma x_t')x_t^\top \\ \lambda x_t^\top \end{bmatrix}, \\ G_{3,t} &\doteq \begin{bmatrix} x_t(x_t^\top - \gamma x_t'^\top) & -\lambda x_t \end{bmatrix}, G_{4,t} \doteq -\xi I, \\ G_{5,t} &\doteq \begin{bmatrix} (1-\gamma)x_{0,t} \\ -\lambda \end{bmatrix}, \end{aligned}$$

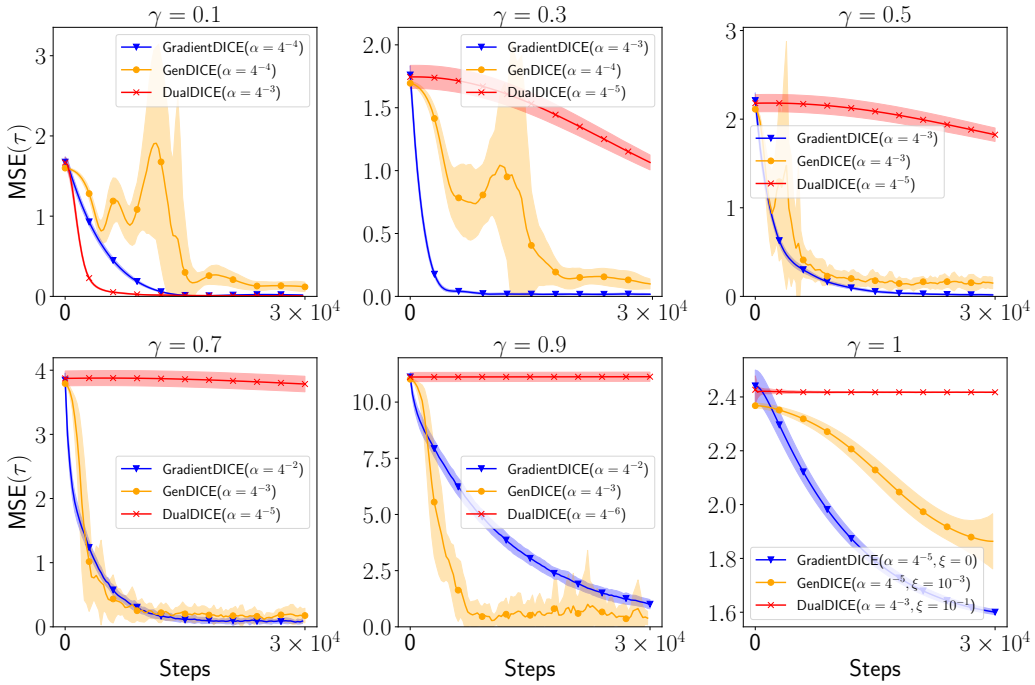

Figure 2. Density ratio learning in Boyan's Chain with a tabular representation.

Algorithm 1 Projected GradientDICE

for $t = 0, \ldots, n-1$ **do**
$y_{t+1} \leftarrow \Pi_Y[y_t + \alpha_t(G_{1,t}y_t + G_{2,t}w_t + G_{5,t})]$
$w_{t+1} \leftarrow \Pi_W[w_t + \alpha_t(G_{3,t}y_t + G_{4,t}w_t)]$
end for
Output:
$\bar{y}_n \doteq \frac{\sum_{t=1}^n \alpha_t y_t}{\sum_{t=1}^n \alpha_t}, \bar{w}_n \doteq \frac{\sum_{t=1}^n \alpha_t w_t}{\sum_{t=1}^n \alpha_t}$

$Y \subset \mathbb{R}^{K+1}, W \subset \mathbb{R}^K$. Π_Y and Π_W are projections onto Y and W w.r.t. ℓ_2 norm (such projection is trivial if Y and W are balls), n is the number of iterations, and α_t is a learning rate, detailed below. We consider the following problem

$$\min_{w \in W} \max_{y \in Y} \left(L(w, y) \doteq L(\tau_w, \eta, f_\kappa) + \frac{\xi}{2}||w||^2\right). \quad (7)$$

It is easy to see $L(w, y)$ is a convex-concave function and its saddle point (w^*, y^*) is unique. We assume

Assumption 4. *Y and W are bounded, closed, and convex, $w^* \in W, y^* \in Y$.*

For the CCSP problem (7), we define the optimization error

$$\epsilon_{\text{opt}}(w, y) \doteq \max_{y' \in Y} L(w, y') - \min_{w' \in W} L(w', y).$$

It is easy to see $L(w, y) = 0$ iif $(w, y) = (w^*, y^*)$.

Proposition 2. *Under Assumptions (1-4), for the $(\bar{w}_n, \bar{y}_n)$ from the Projected GradientDICE algorithm after n iterations, we have at least with probability $1 - \delta$*

$$\epsilon_{opt}(\bar{w}_n, \bar{y}_n) \leq \sqrt{\frac{5}{n}}(8 + 2\ln\tfrac{2}{\delta})C_0,$$

where $C_0 > 0$ is a constant.

Both C_0 and the learning rates α_t are detailed in the proof in the appendix. Note it is possible to conduct a finite sample analysis without introducing projection using arguments from Lakshminarayanan & Szepesvari (2018), which we leave for future work.

5. Experiments

In this section, we present experiments comparing GradientDICE to GenDICE and DualDICE. All curves are averaged over 30 independent runs and shaded regions indicate one standard derivation. The implementations are made publicly available for future research.[4]

5.1. Density Ratio Learning

We consider two variants of Boyan's Chain (Boyan, 1999) as shown in Figure 4. In particular, we use `Episodic Boyan's Chain` when $\gamma < 1$ and `Continuing Boyan's Chain` when $\gamma = 1$. We consider a uniform sampling distribution, i.e., $d_\mu(s, a) = \frac{1}{26} \forall (s, a)$, and a target policy π satisfying $\pi(a_0|s) = 0.1 \forall s$. We design

[4] https://github.com/ShangtongZhang/DeepRL

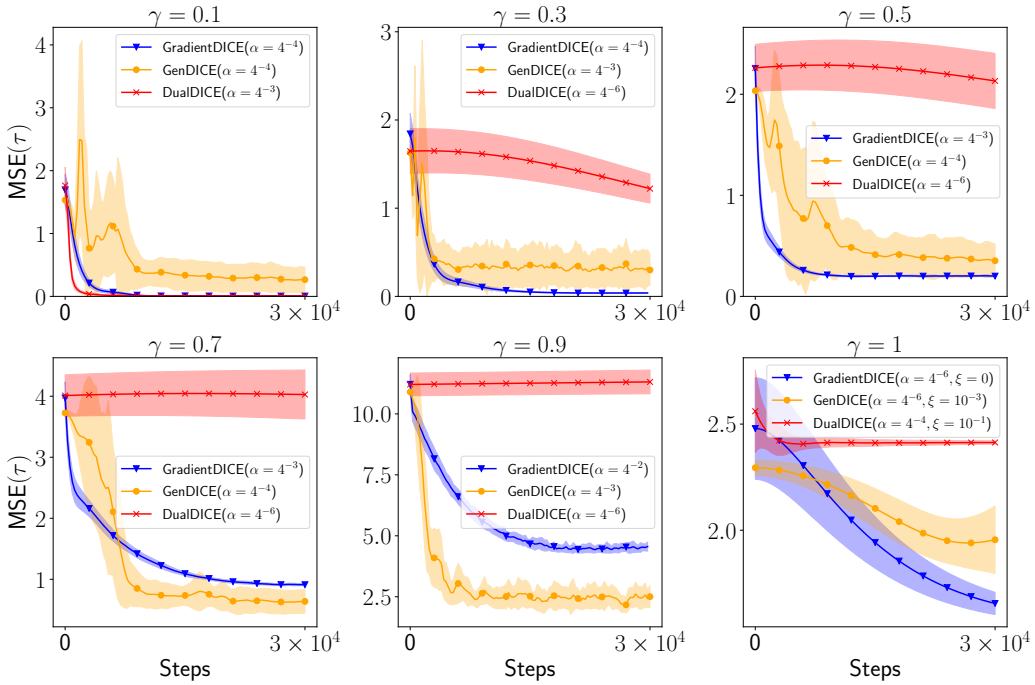

Figure 3. Density ratio learning in Boyan's Chain with a linear architecture.

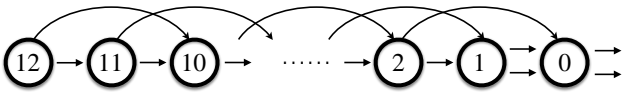

Figure 4. Two variants of Boyan's Chain. There are 13 states in total with two actions $\{a_0, a_1\}$ available at each state. The initial distribution $\tilde{\mu}_0$ is uniform over $\{s_0, \ldots, s_{12}\}$. At a state $s_i (i \geq 2)$, a_0 leads to s_{i-1} and a_1 leads to s_{i-2}. At s_1, both actions leads to s_0. At s_0, there are two variants. (1) `Episodic Boyan's Chain`: both actions at s_0 lead to s_0 itself, i.e., s_0 is an absorbing state. (2) `Continuing Boyan's Chain`: both actions at s_0 lead to a random state among $\{s_0, \ldots, s_{12}\}$ with equal probability.

a sequence of tasks by varying the discount factor γ in $\{0.1, 0.3, 0.5, 0.7, 0.9, 1\}$.

We train all compared algorithms for 3×10^4 steps. We evaluate the Mean Squared Error (MSE) for the predicted τ every 300 steps, computed as $\text{MSE}(\tau) \doteq \frac{1}{26} \sum_{s,a} (\tau(s,a) - \tau_*(s,a))^2$, where the ground truth τ_* is computed analytically. We use fixed learning rates α for all algorithms, which is tuned from $\{4^{-6}, 4^{-5}, \ldots, 4^{-1}\}$ to minimize the MSE(τ) at the end of training. For the setting $\gamma = 1$, we additionally tune ξ from $\{0, 10^{-3}, 10^{-2}, 10^{-1}\}$ (for a fair comparison, we also add this ridge regularization for GenDICE and DualDICE). For the penalty coefficient, we set $\lambda = 1$ as recommended by Zhang et al. (2020a). We find λ has little influence on the learning process in this domain.

We report the results in both tabular (Figure 2) and linear (Figure 3) settings. In the tabular setting, we use lookup tables to store τ, f and η. In the linear setting, we use two independent sets of weights for the two actions. As GenDICE requires $\tau \succeq 0$, we use the nonlinearity $(\cdot)^2$ for its τ prediction as suggested by Zhang et al. (2020a). We do not apply any nonlinearity for GradientDICE and DualDICE. Our results show that GradientDICE reaches a lower prediction error at the end of training than GenDICE in 4 (5) out 6 tasks in the tabular (linear) setting. Moreover, the learning curves of GradientDICE are more stable than those of GenDICE in all the 6 tasks in both tabular and linear settings. Although DualDICE performs the best for the task $\gamma = 0.1$, it becomes unstable as γ increases, which is also observed in Zhang et al. (2020a).

5.2. Off-Policy Evaluation

We now benchmark DualDICE, GenDICE, and GradientDICE in an off-policy evaluation problem. We consider `Reacher-v2` from OpenAI Gym (Brockman et al., 2016). We consider policies in the form of $\pi_d(s,a) + \mathcal{N}(0, \sigma^2)$, where π_d is a deterministic policy trained via TD3 (Fujimoto et al., 2018) for 10^6 steps and $\mathcal{N}$ is Gaussian noise. For the behavior policy, we set $\sigma = 0.1$ and run the policy for 10^5 steps to collect transitions, which form the dataset used across all the experiments. For the target policy, we set $\sigma = 0.05$.

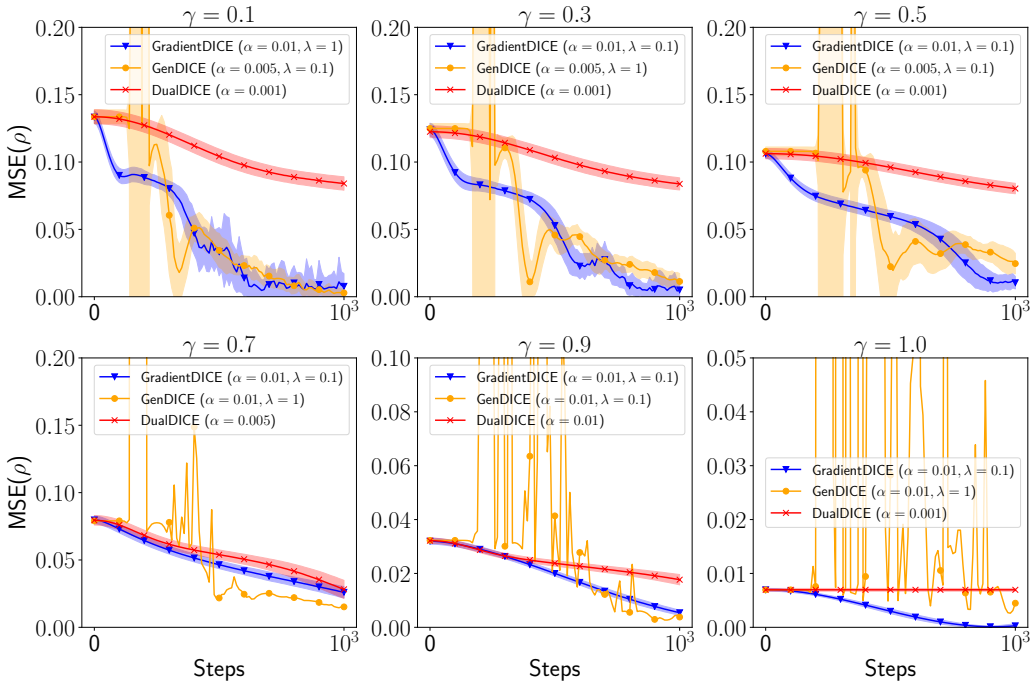

Figure 5. Off-policy evaluation in `Reacher-v2` with neural network function approximators.

We use neural networks to parameterize τ and f, each of which is represented by a two-hidden-layer network with 64 hidden units and ReLU (Nair & Hinton, 2010) activation function. For GenDICE, we add $(\cdot)^2$ nonlinearity for the τ prediction by the network. For GradientDICE and DualDICE, we do not have such nonlinearity in their τ prediction. Given the learned τ, the performance $\rho_\gamma(\pi)$ is approximated by $\hat{\rho}_\gamma(\pi) \doteq \frac{1}{N}\sum_{i=1}^{N} \tau(s_i, a_i) r_i$. We train each algorithm for 10^3 steps and examine $\text{MSE}(\rho) \doteq \frac{1}{2}(\rho_\gamma(\pi) - \hat{\rho}_\gamma(\pi))^2$ every 10 steps, where the ground truth $\rho_\gamma(\pi)$ is computed from Monte Carlo methods via executing the target policy π multiple times. We use SGD to train the neural networks with batch size 128. The learning rate α and the penalty coefficient λ are tuned from $\{0.01, 0.005, 0.001\}$ and $\{0.1, 1\}$ with grid search to minimize MVE(ρ) at the end of training.

The results are reported in Figure 5. Although policy evaluation errors of GradientDICE and GenDICE tend to be similar at the end of training, the learning curves of GradientDICE are more stable than those of GenDICE, which matches the results in the tabular and linear settings. Although DualDICE tends to be more stable than both GradientDICE and GenDICE, it learns slower and does not work for the setting $\gamma = 1$, which also matches the results in Zhang et al. (2020a). To summarize, GradientDICE combines the advantages of both DualDICE (stability under the total discounted reward criterion) and GenDICE (compatibility with the average reward criterion).

6. Related Work

Inspired by Hallak & Mannor (2017); Liu et al. (2018), Gelada & Bellemare (2019); Uehara & Jiang (2019) propose different estimators for learning density ratios, either with semi-gradient updates (Sutton, 1988) or by restricting the function class to Reproducing Kernel Hilbert Space. Zhang et al. (2020b) show that some density ratios can be interpreted as special emphasis (Sutton et al., 2016). Hence the Gradient Emphasis Learning algorithm in Zhang et al. (2020b) can also be used to learn certain density ratios. All these methods, however, require a single known behavior policy. By contrast, DualDICE, GenDICE, and GradientDICE cope well with multiple unknown behavior policies.

7. Conclusion

In this paper, we point out two problems with GenDICE and fix them with GradientDICE. We provide a comprehensive theoretical analysis for GradientDICE. Our experiments confirm that the theoretical advantages of GradientDICE over GenDICE translate into an empirical performance boost in the tested domains. Overall, our work provides a new theoretical justification for the field of density-ratio-learning-based off-policy evaluation.

Acknowledgments

SZ is generously funded by the Engineering and Physical Sciences Research Council (EPSRC). This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement number 637713). The experiments were made possible by a generous equipment grant from NVIDIA. BL's research is funded by the National Science Foundation (NSF) under grant NSF IIS1910794 and Amazon Research Award. The authors thank Lihong Li and Bo Dai for the useful discussion.

References

Baird, L. Residual algorithms: Reinforcement learning with function approximation. *Machine Learning*, 1995.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Boyan, J. A. Least-squares temporal difference learning. In *Proceedings of the 16th International Conference on Machine Learning*, 1999.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

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

Dai, B., He, N., Pan, Y., Boots, B., and Song, L. Learning from conditional distributions via dual embeddings. *arXiv preprint arXiv:1607.04579*, 2016.

Fujimoto, S., van Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*, 2018.

Gelada, C. and Bellemare, M. G. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019.

Hallak, A. and Mannor, S. Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Horn, R. A. and Johnson, C. R. *Matrix analysis (2nd Edition)*. Cambridge university press, 2012.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.

Lakshminarayanan, C. and Szepesvari, C. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *The 21st International Conference on Artificial Intelligence and Statistics*, 2018.

Lin, L.-J. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 1992.

Liu, B., Liu, J., Ghavamzadeh, M., Mahadevan, S., and Petrik, M. Finite-sample analysis of proximal gradient td algorithms. In *UAI*, 2015.

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

Maei, H. R. *Gradient temporal-difference learning algorithms*. PhD thesis, University of Alberta, 2011.

Mousavi, A., Li, L., Liu, Q., and Zhou, D. Black-box off-policy estimation for infinite-horizon reinforcement learning. In *International Conference on Learning Representations*, 2020.

Nachum, O., Chow, Y., Dai, B., and Li, L. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *arXiv preprint arXiv:1906.04733*, 2019.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 2009.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pp. 271–279, 2016.

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

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

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

Sherman, J. and Morrison, W. J. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 1950.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine Learning*, 1988.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction (2nd Edition)*. MIT press, 2018.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning*, 2009a.

Sutton, R. S., Maei, H. R., and Szepesvári, C. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 2009b.

Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., and Precup, D. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems*, 2011.

Sutton, R. S., Mahmood, A. R., and White, M. An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 2016.

Tsitsiklis, J. N. and Van Roy, B. Analysis of temporal-diffference learning with function approximation. In *Advances in Neural Information Processing Systems*, 1997.

Uehara, M. and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

Wang, T., Bowling, M., and Schuurmans, D. Dual representations for dynamic programming and reinforcement learning. In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, 2007.

Wang, T., Bowling, M., Schuurmans, D., and Lizotte, D. J. Stable dual dynamic programming. In *Advances in neural information processing systems*, 2008.

Zhang, R., Dai, B., Li, L., and Schuurmans, D. Gendice: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2020a.

Zhang, S., Liu, B., Yao, H., and Whiteson, S. Provably convergent two-timescale off-policy actor-critic with function approximation. In *Proceedings of the 37th International Conference on Machine Learning*, 2020b.