Imitation from Arbitrary Experience: A Dual Unification of Reinforcement and Imitation Learning Methods

Anonymous Author(s)
Affiliation
Address
email

Abstract

It is well known that Reinforcement Learning (RL) can be formulated as a convex program with linear constraints. The dual form of this formulation is unconstrained, which we refer to as dual RL, and can leverage preexisting tools from convex optimization to improve the learning performance of RL agents. We show that several state-of-the-art deep RL algorithms (in online, offline, and imitation settings) can be viewed as dual RL approaches in a unified framework. This unification calls for the methods to be studied on common ground, so as to identify the components that actually contribute to the success of these methods. Our unification also reveals that prior off-policy imitation learning methods in the dual space are based on an unrealistic coverage assumption and are restricted to matching a particular f-divergence. We propose a new method using a simple modification to the dual framework that allows for imitation learning with arbitrary off-policy data to obtain near-expert performance.

1 Introduction

Most of the modern deep Reinforcement Learning (RL) algorithms are built on the foundations of the approximate dynamic programming (ADP) formulation [2]. This formulation led to a family of temporal difference algorithms and to some of the most performant RL algorithms. Model-free deep RL algorithms such as SAC [17], TD3 [12], DDPG [36] and model-based deep RL algorithms such as Dreamer [18], TD-MPC [19], LOOP [48], MuZero [44] and MBPO [24] exemplify this class of successful algorithms. These methods have accomplished various feats including solving challenging games such as Starcraft, Chess, Go, and Shogi while also being very successful in robotics for learning manipulation and locomotion policies.

In this work, we consider an alternate formulation for RL, which views RL as a convex objective with linear constraints. This framework has been formalized before [40] and has been known in its linear programming form since the work of [38, 5]. The convex program for RL can be converted into a dual unconstrained form that admits the tools already present in convex optimization literature, making it more suitable for stochastic optimization. The methods arising from the convex program formulation of RL also have the benefit of being truly off-policy: they can use arbitrary off-policy data to better estimate the on-policy policy gradient, i.e they can implicitly perform distribution correction. For this reason, the methods in this space have often been referred to as DICE (DIstribution Correction Estimation) methods in previous literature [41, 31, 35, 37, 56].

Our first contribution is a demonstration that several recent algorithms in deep RL and imitation learning can be viewed as dual RL methods in a unified framework. These algorithms represent state-of-the-art methods [15, 57, 34, 14] in reinforcement and imitation learning, for both online and offline settings. A number of these recent works have utilized differing tools to derive their method (e.g Gumbel regression [15], change of variables [14, 57], lower bounded Q-function [34]) which makes it difficult to study them on common ground. We hope that our presented unification provides a framework for future methods to perform evaluation and analyze which factors actually make the algorithm better or worse.

Second, building upon the dual framework, we propose a new algorithm for off-policy imitation learning that is able to leverage arbitrary off-policy data to learn near-expert policies, relaxing the coverage assumption of previous works [37, 7, 57, 29]. Our resulting algorithm, `ReCOIL`, is simple, non-adversarial, and admits a single-player optimization in contrast to previous works in imitation [16, 21, 10, 47]. We empirically demonstrate the failure of previous imitation learning methods based on the coverage assumption in the dual setting. We also evaluate our methods for learning to imitate in offline continuous control settings on a set of MuJoCo environments and show competitive performance. Finally, we point out scope for future exploration in designing better algorithms for deep RL and imitation learning methods utilizing the dual framework.

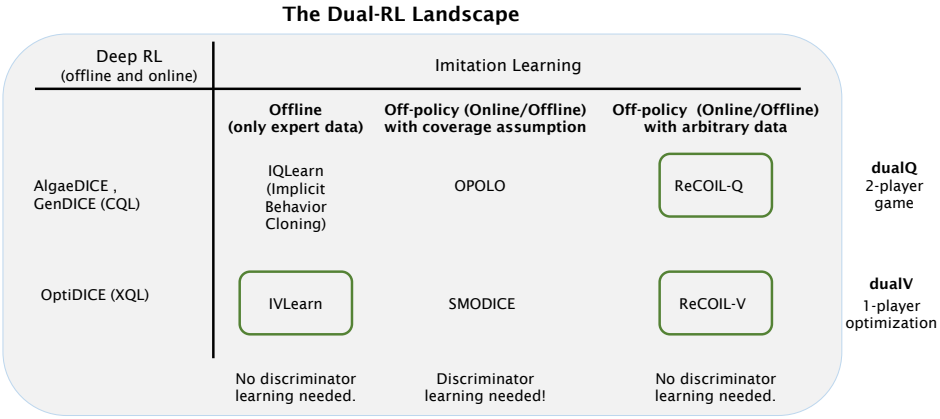

Figure 1: We show that a number of prior methods are special cases of the dual RL framework. Based on this framework, we also propose new methods addressing the shortcomings of previous works (boxed in green).

2 Preliminaries

A Markov Decision Process (MDP) is defined by the tuple $(\mathcal{S}, \mathcal{A}, p, r, d_0)$ with state-space $\mathcal{S}$, action-space $\mathcal{A}$, transition probability $p(s_{t+1}|s_t, a_t)$, reward function $r(s, a)$, and initial state distribution $d_0(s)$. In the infinite horizon discounted MDP, the goal of reinforcement learning algorithms is to maximize the return for policy π given by $J^\pi = \mathbb{E}_{a_t \sim \pi(s_t), s_0 \sim d_0}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$.

Value functions: $V^\pi : \mathcal{S} \to \mathbb{R}$ represents a state-value function which estimates the return from the current state s_t and following policy π, defined as $V^\pi(s) = \mathbb{E}_{a_t \sim \pi(s_t)}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s]$. Similarly, $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ represents an action-value function, usually referred as a Q-function, defined as $Q^\pi(s, a) = \mathbb{E}_{a_t \sim \pi(s_t)}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s, a_0 = a]$. Value functions corresponding to the optimal policy π^* are defined to be V^* and Q^*. The value function can be updated by minimizing mean-squared error with the target given by Bellman operator $\mathcal{T}^{\pi_Q}$ in the ADP setting:

$$\mathcal{T}^{\pi_Q} Q(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{s_{t+1} \sim p, a_{t+1} \sim \pi_Q}[\gamma(Q(s_{t+1}, a_{t+1})], \tag{1}$$

where π_Q is updated to be greedy with respect to Q, the current Q-function. We will use $\mathcal{T}V(s, a)$ to denote $r(s, a) + \mathbb{E}_{s_{t+1} \sim p(\cdot|s_T, a_t)}[V(s_{t+1})]$. Note the absence of an explicit policy in this operator.

State-action/State visitation distribution: The state-action visitation distribution (stationary occupancies) $d^\pi(s,a) : \mathcal{S} \times \mathcal{A} \to [0, \infty)$ of π is:

$$d^\pi(s,a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t P(s_t = s, a_t = a | s_0 \sim \rho_0, a_t \sim \pi(s_t), s_{t+1} \sim p(s_t, a_t)). \tag{2}$$

The state-visitation distribution marginalizes over actions on $d^\pi(s,a)$ and will be overloaded by the notation $d^\pi(s) = \sum_{a \in \mathcal{A}} d^\pi(s,a)$. In this work, we will use d^O, d^R, and d^E to denote offline, replay-buffer, and expert state-action visitation distribution respectively.

f-divergence: Let $f : (0, \infty) \to \mathbb{R}$ be a convex lower semi-continuous function with $f(1) = 0$. Let P and Q be two probability distributions, then the f-divergence is defined as:

$$D_{\mathrm{f}}(P \,||\, Q) = \mathbb{E}_{z \sim Q}\left[f\left(\frac{P(z)}{Q(z)} \right) \right]. \tag{3}$$

3 A Dual Unification for Reinforcement Learning Methods

A number of recently proposed algorithms for deep reinforcement learning and imitation learning use different mathematical tools to derive their method (e.g Gumbel regression [15], change of variables [14, 57], lower-bounded Q-function [34]) which makes it difficult to study them on common ground. This paper aims to connect a number of algorithmic developments in deep reinforcement learning as a special case in a simple unified framework of dual reinforcement learning. (True) off-policy algorithms, that can leverage arbitrary off-policy data for policy improvement, are a simple by-product of this previously known but mostly ignored in-practice framework [40] we consider. In Section 3.1, we first describe the general framework for dual reinforcement learning and then proceed to chalk out the connections for reinforcement learning and imitation learning in Section 3.2 and Section 3.3 respectively. In Section 4, we proceed to propose a new method, `ReCOIL`, for imitation learning from arbitrary experience.

3.1 Dual Reinforcement Learning

We will consider the regularized policy optimization setting in this work given below:

$$\max_{\pi} \mathbb{E}_{d^\pi(s,a)}[r(s,a)] - \alpha D_{\mathrm{f}}(d^\pi(s,a) \,||\, d^O(s,a)), \tag{4}$$

where d^O is a known state-action visitation distribution, α is the temperature parameter that allows us to weigh policy improvement against conservatism by staying close to the state-action distribution d^O, and f denotes a particular f-divergence. We can rewrite the problem as a convex optimization problem (CoP) by considering optimization over *valid* state-action visitations satisfying bellman-flow constraints:

$$\textbf{Q-CoP:} \quad \max_{\pi, d \geq 0} \mathbb{E}_{d(s,a)}[r(s,a)] - \alpha D_{\mathrm{f}}(d(s,a) \,||\, d^O(s,a)) \tag{5}$$
$$\text{s.t } d(s,a) = (1-\gamma) d_0(s).\pi(a|s) + \gamma \sum_{s',a'} d(s',a') p(s|s',a') \pi(a|s),$$

The above problem is overconstrained – the inner maximization w.r.t d is unnecessary as the $(\mathcal{S} \times \mathcal{A})$ constraints uniquely determine the distribution. We can relax the constraints to get an equivalent but different form for the problem:

$$\textbf{V-CoP:} \quad \max_{d \geq 0} \mathbb{E}_{d(s,a)}[r(s,a)] - \alpha D_{\mathrm{f}}(d(s,a) \,||\, d^O(s,a)) \tag{6}$$
$$\text{s.t } \sum_{a \in \mathcal{A}} d(s,a) = (1-\gamma) d_0(s) + \gamma \sum_{s',a'} d(s',a') p(s|s',a').$$

We can leverage Lagrangian duality to convert the convex program above into its dual unconstrained optimization form (see Appendix A.1 for derivation and review) since the objective is convex and the constraints are linear.

In summary, we have two methods for policy optimization given by:

$$\texttt{dual-Q}: \max_{\pi} \min_{Q} (1-\gamma)\mathbb{E}_{d_0(s),\pi(a|s)}[Q(s,a)] + \alpha\mathbb{E}_{s,a\sim d^O}\left[f^*\left(\left[\mathcal{T}^\pi Q(s,a) - Q(s,a)\right]/\alpha\right)\right] \tag{7}$$

$$\texttt{dual-V}: \min_{V(s)} (1-\gamma)\mathbb{E}_{d_0(s)}[V(s)] + \alpha\mathbb{E}_{s,a\sim d^O}\left[f^*\left(\left[\mathcal{T}V(s,a) - V(s)\right]/\alpha\right)\right]. \tag{8}$$

where Q and V are Lagrange variables in this framework and f^* denotes the convex conjugate of function f. We overload $y(s,a)$ to denote $r(s,a) + \gamma \sum_{s'} p(s'|s,a)\pi(a'|s')Q(s',a') - Q(s,a)$ or $r(s,a) + \gamma \sum_{s'} p(s'|s,a)V(s') - V(s)$ for `dual-Q` and `dual-V` settings respectively for notational simplicity. We note an important property of the dual RL formulation: at convergence, the distribution ratio between the optimal policy and distribution used for regularization is captured in the first derivative of the conjugate f-divergence. Formally,

$$\frac{d^*(s,a)}{d^O(s,a)} = {f^*}'\left(\frac{y(s,a)}{\alpha}\right). \tag{9}$$

The dual formulation has two appealing properties: (a) The gradient of the objective (Eq 7 and 8) w.r.t π as Q is optimized is the on-policy policy gradient using off-policy data[1] (b) Allows us to incorporate regularization of different forms such as w.r.t behavior data (pessimism [53]), data generated by previous policy (trust region [26, 45, 46]), uniform distribution (max-entropy [17]) or an online planner [48]. In the dual formulation above, we have ignored the constraints that imply that the distribution $d(s,a)$ should be always positive $\forall\ s,a$. The closed-form solutions for the dual formulation under the positivity constraint can be found in Appendix A.2. The positivity constraints have the effect of changing the function f^* in Equation 7 and 8 to f_p^* (refer Appendix A.2).

3.2 Connections to duality in reinforcement learning

Equations 5 and 6 provide a natural choice for both offline and online reinforcement learning algorithms. Using the dual RL framework, we obtain a set of approaches that leverage `dual-Q` and `dual-V` for policy improvement. In this section, our first result shows that Conservative Q-learning [34], an offline RL method primarily understood to prevent overestimation by learning a lower bounded Q function is actually a `dual-Q` method. Second, we show that Extreme Q-Learning (X-QL) [15], a method for both online and offline RL based on the principle of *implicit maximization* in the value function space using Gumbel regression, can be reduced to a `dual-V` problem with a semi-gradient update rule (i.e `stop-gradient` $(r(s,a) + \gamma \sum_{s'} p(s'|s,a)V(s'))$) when f is set to be the reverse-KL regularization. This insight obtained through duality, allows us to propose a class of algorithms extending X-QL, by choosing different functions f which we show below to result in a family of implicit maximizers. We defer the detailed discussion on unifying RL methods using duality to Appendix A.3 with Lemmas 5, 6, 7 and focus on the imitation learning setting in the main paper due to space constraints.

3.3 Connections to duality in imitation learning

In this section, we will first discuss the setting of offline imitation learning with expert data only – where the agent has no access to the environment and is limited to a fixed amount of offline expert transitions. In this direction, we consolidate prior work and also propose a new algorithm arising from the dual-V formulation. Then, we discuss the off-policy imitation learning setting where the agent has access to limited expert data that it is trying to imitate along with some suboptimal data which is obtained offline or through agent online interactions. We discuss prior works from the perspective of duality and show how they are limited by their assumption of coverage (off-policy replay data covers expert data) or their reliance on a particular f-divergence. Finally, we propose a new method for off-policy imitation learning that relaxes the coverage assumption and works for arbitrary f-divergences.

[1] Note that the on-policy policy gradient is with respect to the regularized Q-values

3.3.1 Offline imitation learning with expert data only

We start with the `dual-Q` and `dual-V` equations and repurpose them for imitation by simply setting the reward function to be uniformly 0 across the state-action space and setting the regularization distribution to be the expert distribution $d^O(s,a) = d^E(s,a)$. We use $\mathcal{T}_0^\pi$ and $\mathcal{T}_0$ to denote the backup operator with zero rewards. `dual-Q` (offline imitation) takes the following form:

`dual-Q` (offline imitation with expert-only data):

$$\max_{\pi} \min_{Q} (1-\gamma)\mathbb{E}_{d_0(s),\pi(a|s)}[Q(s,a)] + \mathbb{E}_{s,a\sim d^E}[f^*\left([\mathcal{T}_0^\pi Q(s,a) - Q(s,a)]\,/\alpha\right)]. \tag{10}$$

Interestingly, this reduction directly leads us to the equivalence of an imitation learning method IQLearn [14] derived using a change of variables in the form of an inverse backup operator. [14] uses this method in the online imitation learning setting with an additional regularization which we suggest is unprincipled, also pointed out by others [1] (as only expert data samples can be leveraged in the above optimization) and provide a fix in the Section 4.

Lemma 1. *`dual-Q` is equivalent to IQ-Learn when $r(s,a) = 0 \;\; \forall\, (\mathcal{S},\mathcal{A})$ and $d^O(s,a) = d^E(s,a)$.*

Utilizing Lemma 1 above, we also provide a new observation in the form of Corollary 1 below. We find that Implicit Behavior Cloning [7], a method that performs behavior cloning using a contrastive objective, is a special case of `dual-Q` for offline imitation learning.

Corollary 1. *`dual-Q` is equivalent to Implicit Behavior Cloning [7] when $r(s,a) = 0 \;\; \forall\, (\mathcal{S},\mathcal{A})$ and $d^O(s,a) = d^E(s,a)$ and f is set to be the total variation divergence.*

Leveraging the `dual-V` variant we analogously obtain a new method for offline imitation learning, which we refer to as `IVLearn` that we leave for future work. (See Appendix A.4.1).

3.3.2 Off-policy imitation learning (under coverage assumption)

The `dual-Q` and `dual-V` framework does not naturally extend to off-policy imitation learning. To remedy this, prior methods have relied on careful selection of an f-divergence and a coverage assumption that allows them to arrive at an off-policy objective for imitation learning [57, 22, 37]. *Under the assumption that the replay buffer visitation (denoted by d^R) covers the expert visitation ($d^R > 0$ wherever d^E>0)* [37], which we refer to as the **coverage assumption**, and f set to be the reverse KL divergence, we get the following `dual-Q` and `dual-V` forms for imitation (see Appendix A.5 for derivation):

`dual-Q` for off-policy imitation (coverage assumption):

$$\max_{\pi(a|s)} \min_{Q(s,a)} (1-\gamma)\mathbb{E}_{\rho_0(s),\pi(a|s)}[Q(s,a)] + \mathbb{E}_{s,a\sim d^R}[f^*(\mathcal{T}^\pi_{r^{imit}} Q(s,a) - Q(s,a))], \tag{11}$$

where $\mathcal{T}^\pi_{r^{imit}}$ denote backup operator under the reward function $r^{imit} = -\log \frac{d^R(s,a)}{d^E(s,a)}$. This choice of KL divergence leads us to a reduction of another method, OPOLO [57] for off-policy imitation learning to `dualQ` which we formalize in the lemma below:

Lemma 2. *`dual-Q` for off-policy imitation learning reduces to OPOLO [57], with the f-divergence set to the reverse KL divergence when $r(s,a) = 0\; \forall \mathcal{S},\mathcal{A}$, $d^O = d^E$ and under the assumption that the replay data distribution covers the expert data distribution.*

We note that the `dual-V` framework for off-policy imitation learning under coverage assumptions was studied in the imitation learning work SMODICE [37].

4 A Method for Imitation Learning from Arbitrary Experience

In this section, we propose a new approach to relax the coverage assumption discussed above for imitation as well as allow for arbitrary f-divergence matching in the dual framework. We refer to our method as `ReCOIL` (RElaxed Coverage for Off-policy Imitation Learning). We consider an

alternate optimization objective that fits well in the dual RL framework by changing the regularization to be between mixture distributions. Concretely, we are interested in the following f-divergence regularization:

$$D_{\mathrm{f}}(\beta d(s,a) + (1-\beta)d^R(s,a) \,||\, \beta d^E(s,a) + (1-\beta)d^R(s,a)). \tag{12}$$

Minimizing this divergence alone is a valid imitation learning objective [16, 27, 43, 47] since the global minimum for this objective is achieved at $d = d^E$. Let $d^R_{mix} := \beta d(s,a) + (1-\beta)d^R(s,a)$ and $d^{E,R}_{mix} := \beta d^E(s,a) + (1-\beta)d^R(s,a)$. The modified Q-CoP for imitation learning objective is given by:

$$\begin{aligned} &\max_{d(s,a)\geq 0, \pi(a|s)} -D_{\mathrm{f}}(d^R_{mix}(s,a) \,||\, d^{E,R}_{mix}(s,a)) \\ &\text{s.t } d(s,a) = (1-\gamma)\rho_0(s).\pi(a|s) + \gamma\pi(a|s)\sum_{s',a'} d(s',a')p(s|s',a') \end{aligned} \tag{13}$$

The V-CoP for this setting can be similarly specified and the dual results in two variants of ReCOIL that can be leveraged for off-policy imitation learning with arbitrary data as formalized by Lemma 3 and Lemma 4 below.

Lemma 3. ***(`dual-Q` for off-policy imitation (relaxed coverage assumption))*** *Imitation learning using off-policy data can be solved by optimizing the following modified dual objective for Q-CoP with $r(s,a) = 0\ \forall \mathcal{S}, \mathcal{A}$ and f-divergence considered between distributions $d^R_{mix} := \beta d(s,a) + (1-\beta)d^R(s,a)$ and $d^{E,R}_{mix} := \beta d^E(s,a) + (1-\beta)d^R(s,a)$, and is given by:*

$$\begin{aligned} \max_{\pi(a|s)} \min_{Q(s,a)} &\beta(1-\gamma)\mathbb{E}_{d_0(s),\pi(a|s)}[Q(s,a)] + \mathbb{E}_{s,a\sim d^{E,R}_{mix}}\left[f^*_p(\mathcal{T}^\pi_0 Q(s,a) - Q(s,a))\right] \\ &- (1-\beta)\mathbb{E}_{s,a\sim d^R}[\mathcal{T}^\pi_0 Q(s,a) - Q(s,a)] \end{aligned} \tag{14}$$

Analogously, we have the following result for off-policy imitation learning in the V-space.

Lemma 4. ***(`dual-V` for off-policy imitation (relaxed coverage assumption))*** *Imitation learning using off-policy data can be solved by optimizing the following modified dual objective for V-CoP with $r(s,a) = 0\ \forall \mathcal{S}, \mathcal{A}$ and f-divergence considered between distributions $d^R_{mix} := \beta d(s,a) + (1-\beta)d^R(s,a)$ and $d^{E,R}_{mix} := \beta d^E(s,a) + (1-\beta)d^R(s,a)$, and is given by:*

$$\begin{aligned} \min_{V(s)} &\beta(1-\gamma)\mathbb{E}_{d_0(s)}[V(s)] + \mathbb{E}_{s,a\sim d^{E,R}_{mix}}\left[f^*_p(\mathcal{T}_0 V(s,a) - V(s))\right] \\ &- (1-\beta)\mathbb{E}_{s,a\sim d^R}[\mathcal{T}_0 V(s,a) - V(s)] \end{aligned} \tag{15}$$

In the above-proposed methods, we avoid the pitfalls of previous methods based on coverage assumption that requires learning a reward function $\left(\log(\frac{d^R(s,a)}{d^E(s,a)})\right)$ that is ill-defined in state-action space with zero expert support. Our methods also implicitly learns a distribution ratio $\frac{\beta d(s,a)+(1-\beta)d^R(s,a)}{\beta d^E(s,a)+(1-\beta)d^R(s,a)}$ as $f^{*'}(y(s,a))$ which is well-defined for all replay buffer and expert transitions ($d^R > 0$ or $d^E > 0$) that the policy is trained on.

5 Experimental Analysis

Our experimental evaluation aims to answer the following questions: (a) Does the (true) off-policy nature of dual RL framework allow it to fix the brittleness in traditional off-policy methods? (b) Does `ReCOIL` allow for better estimation of agent visitation distribution? (c) How does `ReCOIL` compare to baselines for learning to imitate from offline datasets of mixed quality?

5.1 The failure of ADP-based traditional off-policy algorithms

Our experiments with the popular off-policy method SAC [17] reveal its brittleness to off-policy data. At the beginning of training, each learning agent is provided with expert or human-demonstrated

trajectories for completing the task. We add 1000 transitions from this dataset to the replay buffer for the off-policy algorithm to bootstrap from. As the action dimension increases, the brittleness of SAC becomes more apparent (see SAC+off policy data and SACfD plots in Figure 3). We hypothesize that this failure in the online RL setting is primarily due to the training instabilities caused by TD-backups resulting in overestimation in regions where the agent's current policy does not visit. In Figure 4, we observe that overestimation indeed happens in environments with larger action dimensions and these overestimations take longer to get corrected and in the process destabilize the training. Figure 3 in Appendix D.1 shows that the `dual-Q` method for RL (AlgaeDICE) is able to leverage off-policy data to increase learning performance without any signs of destabilization. This can be attributed to the distribution correction estimation property of dual RL methods which updates the current policy using the corrected on-policy policy visitation [41]. Note, that we set the temperature α to a low value (0.001) to disentangle the effect of pessimism which is an alternate way to avoid overestimation.

5.2 Does `ReCOIL` allow for better estimation of agent visitation distribution?

We consider the proposed `ReCOIL` method and investigate its ability to estimate distribution ratios correctly using the property shown in Eq 9. We consider the inner optimization for `ReCOIL-Q` to obtain the distribution ratio for a fixed policy π. Thus, given the visitation distribution of the replay buffer d^R, expert d^E, and the policy π, the inner optimization implicitly learn the distribution ratio between d^R_{mix} and $d^{E,R}_{mix}$, allowing us to infer agent visitation d^π. We consider two settings in our experiments: (1) a 2-timestep MDP where the agent starts from state s_0 and transitions to one of the states $\{s_1, s_2, s_3, s_4, s_5\}$ which are absorbing. In this setting, the replay buffer perfectly covers the unknown ground truth agent visitation. (2) a 2-D gridworld where the agent can move cardinally and the replay buffer distribution does not cover the unknown ground truth agent visitation.

Our results (Figure 5 and 6 in Appendix D.2) demonstrate that in the perfect coverage setting, `ReCOIL` is able to infer the agent policy visitation perfectly, and in the case of imperfect coverage is able to significantly outperform other methods. Note that we modify SMODICE to incorporate state-action expert data rather than relying on state-only expert data. IQLearn does not leverage replay data information, relying only on expert data to infer agent visitation. SMODICE's reward function $-\log \frac{d^R(s,a)}{d^E(s,a)}$, arising from its coverage assumption is ill-defined in parts of state space where the expert has no support leading to poor downstream density ratio estimation.

5.3 Benchmarking performance of ReCOIL on MuJoCo tasks

Dataset	Env	RCE	ORIL	SMODICE	ReCOIL
random+expert	hopper	51.41±38.63	73.93±11.06	**101.61±7.69**	95.04±4.48
	halfcheetah	64.19±11.06	60.49±3.53	80.16±7.30	**84.10±4.72**
	walker2d	20.90±26.80	2.86±3.39	**105.86±3.47**	100.84±6.34
	ant	105.38±14.15	73.67±12.69	**126.78±5.12**	**126.74±4.63**
random+few-expert	hopper	25.31±18.97	42.04±13.76	60.11±18.28	**79.44±13.53**
	halfcheetah	2.99±1.07	2.84±5.52	2.28±0.62	**3.90±0.66**
	walker2d	40.49±26.52	3.22±3.29	**107.18±1.87**	83.23±19.00
	ant	67.62±15.81	25.41 ± 8.58	-6.10±7.85	**94.25± 8.30**
medium+expert	hopper	29.37±3.39	61.35±7.91	54.77±6.2	**63.26±4.99**
	halfcheetah	61.14±18.31	57.15±3.49	58.01±3.12	**84.01±3.40**
	walker2d	19.84±2.71	4.14±2.69	1.2±1.57	**91.51±16.50**
	ant	81.18±29.73	103.42±8.43	102.82±4.63	**119.94±4.52**

Table 1: Normalized scores for `ReCOIL` on the different D4RL datasets along with 1000 given expert transitions compared to different offline IL baselines. `ReCOIL` improves over baselines when the offline dataset has low coverage of expert or when it is more challenging to distinguish expert vs suboptimal transitions.

In this task, we use an offline dataset of environment interactions from D4RL [8]. We consider the following MuJoCo environments: Hopper, Walker2d, HalfCheetah, and Ant. We consider three different dataset compositions for the offline dataset - 'random+expert', 'medium+expert', and 'random+few-expert'. The first two datasets consist of a few expert trajectories ($\leq$200) and a large number of suboptimal transitions (1 million). random+few-expert has only ≤ 30 expert trajectories. The learning agent also has access to 1000 expert transitions. We compare recent methods for offline

imitation learning with suboptimal data – RCE [6], SMODICE [37] and ORIL [59]. We do not compare to DEMODICE [29] as SMODICE was shown to be competitive in [37].

We focus on `ReCOIL-V` due to its favorable property of being a single-player non-adversarial optimization. Note, that contrary to SMODICE [37] which is based on the coverage assumption and requires learning a discriminator, `ReCOIL` just learns a single parametric V function. Table 1 empirically validates that `ReCOIL-V` outperforms or is competitive over baselines in all environments. SMODICE shows poor performance in cases when the dataset has low expert coverage (random+few-expert) or where the discriminator is not able to reliably distinguish expert vs non-expert data (medium+expert).

6 Related Work

Developments in deep model-free reinforcement learning methods: The promise of off-policy learning falls short as learning off-policy with function approximators presents a number of issues like over-estimation, training instability, and various biases [51, 9, 12, 33]. Previous works have approached fixing these issues using methods like doubleQ learning [20], target networks [39], emphatic weightings [25, 23] among other approaches. These approaches still do not carry over well to the offline RL setting where overestimation bias is more likely since unlike online RL, the policy cannot correct its overestimation by deploying the current policy. A number of solutions exists for controlling overestimation in prior work – f-divergence regularization to the training distribution [53, 42, 13, 11], support regularization [49], implicit maximization [32], learning a Q function that penalizes OOD actions [34] and learning in a pessimistic MDP [28, 55]. A recent method XQL [15] uses implicit maximization to obtain significant gains in learning performance across online and offline RL settings. Prior work [54, 52, 32, 42] investigates specialized solutions to incorporate previously collected off-policy data during online learning. The dual RL framework [40] fixes the issue of distribution mismatch in ADP-based off-policy reinforcement learning works that ignore the distribution shift [17, 12] and provides a principled solution.

Developments in deep model-free imitation learning methods: Imitation learning has benefited greatly by using off-policy data to improve learning performance [30, 43, 47, 57]. Often, replacing the on-policy expectation common in most Inverse RL formulations [58, 50] by expectation under off-policy samples, which is unprincipled, has led to large gains in sample efficiency [30]. Principled offline imitation learning using expert data only is a simple by-product of the dual-RL framework. Previous works have proposed a solution in the dual RL space to this problem based on an unrealistic coverage assumption [37, 57, 29] and restricting themselves to matching a particular f-divergence. In this work, we relax this assumption and allow for generalizing to all f-divergences, presenting a principled off-policy approach to imitation. Our work also presents an approach that allows for single-player non-adversarial optimization for imitation learning in contrast to previous works [31].

7 Conclusion

Dual reinforcement learning algorithms have great potential for developing performant Deep RL methods. Indeed, we show that a number of the recently developed methods for RL – Extreme Q-learning, Conservative Q-learning, and Imitation learning – OPOLO, IQLearn, and Implicit Behavior Cloning can be unified through the lens of an already existing framework of dual RL. Our insight calls for these methods to be studied under this unified lens to figure out components that contribute to the success of these methods. As a by-product of this unification, we are able to identify novel insights that give us a way to propose: a family of implicit maximizers that avoids the overestimation problem due to actor-critic learning by relying on in-distribution maximization, an offline single-player non-adversarial offline imitation learning method from expert data only, and a general off-policy imitation learning method from arbitrary data that relaxes the restrictive coverage assumption made by prior work. We analyze the off-policy imitation learning method by experiments in continuous control and demonstrate superior performance.

References

[1] Anonymous. LS-IQ: Implicit reward regularization for inverse reinforcement learning. In *Submitted to The Eleventh International Conference on Learning Representations*, 2023. under review.

[2] Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming: an overview. In *Proceedings of 1995 34th IEEE conference on decision and control*, volume 1, pages 560–564. IEEE, 1995.

[3] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] Bo Dai, Albert Shaw, Niao He, Lihong Li, and Le Song. Boosting the actor with dual critic. *arXiv preprint arXiv:1712.10282*, 2017.

[5] Eric V Denardo. On linear programming in a markov decision problem. *Management Science*, 16(5):281–288, 1970.

[6] Ben Eysenbach, Sergey Levine, and Russ R Salakhutdinov. Replacing rewards with examples: Example-based policy search via recursive classification. *Advances in Neural Information Processing Systems*, 34:11541–11552, 2021.

[7] Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on Robot Learning*, pages 158–168. PMLR, 2022.

[8] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[9] Justin Fu, Aviral Kumar, Matthew Soh, and Sergey Levine. Diagnosing bottlenecks in deep q-learning algorithms. In *International Conference on Machine Learning*, pages 2021–2030. PMLR, 2019.

[10] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

[11] Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

[12] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[13] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.

[14] Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. Iq-learn: Inverse soft-q learning for imitation. *Advances in Neural Information Processing Systems*, 34:4028–4039, 2021.

[15] Divyansh Garg, Joey Hejna, Matthieu Geist, and Stefano Ermon. Extreme q-learning: Maxent rl without entropy. *arXiv preprint arXiv:2301.02328*, 2023.

[16] Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, pages 1259–1277. PMLR, 2020.

[17] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[18] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[19] Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955*, 2022.

[20] Hado Hasselt. Double q-learning. *Advances in neural information processing systems*, 23, 2010.

[21] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29:4565–4573, 2016.

[22] Hana Hoshino, Kei Ota, Asako Kanezaki, and Rio Yokota. Opirl: Sample efficient off-policy inverse reinforcement learning via distribution matching. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 448–454. IEEE, 2022.

[23] Ehsan Imani, Eric Graves, and Martha White. An off-policy policy gradient theorem using emphatic weightings. *Advances in Neural Information Processing Systems*, 31, 2018.

[24] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

[25] Ray Jiang, Tom Zahavy, Zhongwen Xu, Adam White, Matteo Hessel, Charles Blundell, and Hado Van Hasselt. Emphatic algorithms for deep reinforcement learning. In *International Conference on Machine Learning*, pages 5023–5033. PMLR, 2021.

[26] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[27] Liyiming Ke, Sanjiban Choudhury, Matt Barnes, Wen Sun, Gilwoo Lee, and Siddhartha Srinivasa. Imitation learning as f-divergence minimization. In *Algorithmic Foundations of Robotics XIV: Proceedings of the Fourteenth Workshop on the Algorithmic Foundations of Robotics 14*, pages 313–329. Springer International Publishing, 2021.

[28] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823, 2020.

[29] Geon-Hyeong Kim, Seokin Seo, Jongmin Lee, Wonseok Jeon, HyeongJoo Hwang, Hongseok Yang, and Kee-Eung Kim. Demodice: Offline imitation learning with supplementary imperfect demonstrations. In *International Conference on Learning Representations*, 2022.

[30] Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. *arXiv preprint arXiv:1809.02925*, 2018.

[31] Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. *arXiv preprint arXiv:1912.05032*, 2019.

[32] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

[33] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

[34] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

[35] Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pages 6120–6130. PMLR, 2021.

[36] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[37] Yecheng Jason Ma, Andrew Shen, Dinesh Jayaraman, and Osbert Bastani. Smodice: Versatile offline imitation learning via state occupancy matching. *arXiv preprint arXiv:2202.02433*, 2022.

[38] Alan S Manne. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960.

[39] 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.

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

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

[42] Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

[43] Tianwei Ni, Harshit Sikchi, Yufei Wang, Tejus Gupta, Lisa Lee, and Ben Eysenbach. f-irl: Inverse reinforcement learning via state marginal matching. In *Conference on Robot Learning*, pages 529–551. PMLR, 2021.

[44] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[45] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

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

[47] Harshit Sikchi, Akanksha Saran, Wonjoon Goo, and Scott Niekum. A ranking game for imitation learning. *arXiv preprint arXiv:2202.03481*, 2022.

[48] Harshit Sikchi, Wenxuan Zhou, and David Held. Learning off-policy with online planning. In *Conference on Robot Learning*, pages 1622–1633. PMLR, 2022.

[49] Anikait Singh, Aviral Kumar, Quan Vuong, Yevgen Chebotar, and Sergey Levine. Offline rl with realistic datasets: Heteroskedasticity and support constraints. *arXiv preprint arXiv:2211.01052*, 2022.

[50] Gokul Swamy, Sanjiban Choudhury, J Andrew Bagnell, and Steven Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pages 10022–10032. PMLR, 2021.

[51] Sebastian Thrun and Anton Schwartz. Issues in using function approximation for reinforcement learning. In *Proceedings of the Fourth Connectionist Models Summer School*, volume 255, page 263. Hillsdale, NJ, 1993.

[52] Ikechukwu Uchendu, Ted Xiao, Yao Lu, Banghua Zhu, Mengyuan Yan, Joséphine Simon, Matthew Bennice, Chuyuan Fu, Cong Ma, Jiantao Jiao, et al. Jump-start reinforcement learning. *arXiv preprint arXiv:2204.02372*, 2022.

[53] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

[54] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

[55] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

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

[57] Zhuangdi Zhu, Kaixiang Lin, Bo Dai, and Jiayu Zhou. Off-policy imitation learning from observations. *Advances in Neural Information Processing Systems*, 33:12402–12413, 2020.

[58] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

[59] Konrad Zolna, Alexander Novikov, Ksenia Konyushkova, Caglar Gulcehre, Ziyu Wang, Yusuf Aytar, Misha Denil, Nando de Freitas, and Scott Reed. Offline learning from demonstrations and unlabeled experience. *arXiv preprint arXiv:2011.13885*, 2020.