ALGAE: Policy Gradient from Arbitrary Experience

Ofir Nachum* **Bo Dai*** **Ilya Kostrikov** **Dale Schuurmans**
Google Research
{ofirnachum,bodai,kostrikov,schuurmans}@google.com

Abstract

In many real-world applications of reinforcement learning (RL), interactions with the environment are limited due to cost or feasibility. This presents a challenge to traditional RL algorithms since the max-return objective involves an expectation over on-policy samples. We introduce a new formulation of max-return optimization that allows the problem to be re-expressed by an expectation over an arbitrary *behavior-agnostic* and *off-policy* data distribution. We first derive this result by considering a regularized version of the *dual* max-return objective before extending our findings to unregularized objectives through the use of a Lagrangian formulation of the linear programming characterization of Q-values. We show that, if auxiliary dual variables of the objective are sufficiently optimized, then the gradient of the off-policy objective is *exactly* the on-policy policy gradient, without any use of importance weighting. In addition to revealing the appealing theoretical properties of this approach, we also show that it can deliver good practical performance.

1 Introduction

The use of model-free reinforcement learning (RL) in conjunction with function approximation has proliferated in recent years, demonstrating successful applications in the fields of robotics (Nachum et al., 2019a) and game playing (Mnih et al., 2013). These demonstrations often rely on *on-policy* access to the environment; *i.e.*, during the learning process agents may collect new experience from the environment, and these interactions are effectively unlimited. By contrast, in many real-world applications of RL, interaction with the environment is costly, hence experience collection during learning is limited, necessitating the use of *off-policy* RL methods. In the most restrictive case, direct interaction with the environment is infeasible and learning is completely *offline*; that is, there is only a fixed dataset of logged experience collected by potentially multiple and possibly unknown behavior policies. This scenario is also known as *batch RL* (Lange et al., 2012) with *behavior-agnostic* data.

The off-policy nature of many practical applications presents a challenge for RL algorithms. For example, policy gradient methods (Sutton et al., 2000) require samples from the on-policy distribution to estimate the direction of maximum increase in expected return. The most straightforward way to reconcile policy gradient with off-policy settings is via importance weighting (Sutton et al., 2016). However, this approach is prone to high variance and instability without appropriate damping (Wang et al., 2016; Schulman et al., 2017). Sometimes the easiest solution to a problem is to just ignore it, which is exactly what has been proposed for off-policy policy gradient (Silver et al., 2014; Degris et al., 2012): the suggestion being to simply compute the objective and its gradients with respect to samples from the off-policy data, ignoring any distribution shift. The justification for this approach is that the maximum return policy will be optimal regardless of the state. However, such a justification is unsound in function approximation settings, where models have limited expressiveness, with potentially disastrous consequences on optimization and convergence (Lu et al., 2018).

Value-based methods provide an alternative that may be more promising for the off-policy setting. In these methods, a value function is learned either as a *critic* to a learned policy (as in actor-critic) or as the maximum return value function itself (as in Q-learning). This approach is based on dynamic programming in tabular settings, which is inherently off-policy and independent of any underlying data distribution. Nevertheless, when using function approximation, the objective is traditionally

*Equal contribution.

Optimization Foundations for Reinforcement Learning Workshop at NeurIPS 2019, Vancouver, Canada.

expressed as an expectation over single-step *Bellman errors*, which re-raises the question, "What should the expectation be?" Theoretical work suggests that the ideal expectation is in fact the on-policy expectation (Sutton et al., 2000; Nachum et al., 2018b; Silver et al., 2014). In practice, this problem is usually ignored, with the same justification as that made for policy gradient methods. It is telling that actor-critic or Q-learning algorithms advertised as off-policy still require large amounts of online interaction with the environment (Haarnoja et al., 2018b; Hessel et al., 2018).

In this work, we present an *ALgorithm for policy Gradient from Arbitrary Experience* (ALGAE) as an alternative to policy gradient and value-based methods. We start with the *dual* formulation of the maximum return objective, which is expressed in terms of normalized state-action occupancies rather than a policy or value function. Traditionally, this objective is considered unattractive, since access to the occupancies either requires an on-policy expectation (similar to policy gradient methods) or learning a function approximator to satisfy single-step constraints (similar to value-based methods). We demonstrate how these problems can be remedied by adding convex regularization and applying a carefully chosen change-of-variables, obtaining a joint objective over a policy and an auxiliary dual function (that can be interpreted as a critic). Crucially, this objective relies only on access to samples from an arbitrary off-policy data distribution, collected by potentially multiple and possibly unknown behavior policies. Unlike traditional actor-critic methods, which use a separate objective for actor and critic, this formulation trains the policy (actor) and dual function (critic) to maximize the *same* objective. Further illuminating the connection to policy gradient methods, we show that if the dual function is sufficiently optimized, the gradient of the proposed objective with respect to the policy parameters is *exactly* the on-policy policy gradient. Our approach avoids any explicit use of importance weights and naturally avoids issues of distribution mismatch. Finally, we provide an alternative derivation, based on a primal-dual form of the max-return RL problem, that extends the previous results to both undiscounted $\gamma = 1$ settings and unregularized max-return objectives.

2 Background

We consider the RL problem presented as a Markov Decision Process (MDP) (Puterman, 1994), which is specified by a tuple $\mathcal{M} = \langle S, A, r, T, \mu_0 \rangle$ 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 by starting at an initial state $s_0 \sim \mu_0$, and iteratively producing a sequence of distributions $\pi(\cdot|s_t)$ over A, at steps $t = 0, 1, ...$, from which actions a_t are sampled and successively applied to the environment. The environment produces a scalar reward $r_t = r(s_t, a_t)$ and a next state $s_{t+1} \sim T(s_t, a_t)$. In RL, one wishes to learn a max-return policy; i.e., one wishes to optimize

$$\max_\pi J_{\mathrm{P}}(\pi) := (1-\gamma)\, \mathbb{E}_{s_0 \sim \mu_0, a_0 \sim \pi(s_0)} \left[Q_\pi(s_0, a_0) \right], \tag{1}$$

where Q_π describes the future rewards accumulated by π from any state-action pair (s, a),

$$Q_\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \;\middle|\; s_0 = s, a_0 = a, s_t \sim T(s_{t-1}, a_{t-1}), a_t \sim \pi(s_t) \text{ for } t \geq 1 \right], \tag{2}$$

and $0 \leq \gamma < 1$ is a *discount* factor. This objective may be equivalently written in its *dual* form (Wang et al., 2008) in terms of the policy's normalized state visitation distribution as

$$\max_\pi J_{\mathrm{D}}(\pi) := \mathbb{E}_{(s,a) \sim d^\pi} \left[r(s,a) \right], \tag{3}$$

where

$$d^\pi(s,a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr\left[s_t = s, a_t = s \,|\, s_0 \sim \mu_0, a_t \sim \pi(s_t), s_{t+1} \sim T(s_t, a_t) \text{ for } t \geq 0 \right]. \tag{4}$$

As discussed in Section 4 and Appendix A, these objectives are the primal and dual of the same linear programming (LP) problem. It is common to regularize the max-return objective to encourage either *conservative* behavior (e.g., relative entropy penalty (Nachum et al., 2017b)), or *exploratory* behavior (e.g., causal entropy regularization (Nachum et al., 2017a)); for example, by re-expressing the problem with a modified reward $\tilde{r}(s,a) = r(s,a) - \alpha \log \pi(a|s)$ in the latter case.

In function approximation settings, optimizing π requires access to gradients. The *policy gradient theorem* (Sutton et al., 2000) provides the gradient of $J_{\mathrm{P}}(\pi)$ as

$$\tfrac{\partial}{\partial \pi} J_{\mathrm{P}}(\pi) = \mathbb{E}_{(s,a) \sim d^\pi} \left[Q_\pi(s,a) \nabla \log \pi(a|s) \right]. \tag{5}$$

To properly estimate this gradient one requires access to *on-policy* samples from d^π and access to estimates of the Q-value function $Q_\pi(s,a)$. The first requirement means that every gradient estimate of J_{P} necessitates interaction with the environment, which limits applicability of this method in settings where interaction with the environment is expensive or infeasible. The second requirement means that one must maintain estimates of the Q-function to learn π. This leads to the family of *actor-*

critic algorithms that alternate between updates to π (the actor) and updates to a Q-approximator Q_θ (the critic). The critic is learned by encouraging it to satisfy single-step Bellman consistencies,

$$Q_\pi(s,a) = \mathcal{B}_\pi Q_\pi(s,a) := r(s,a) + \gamma \cdot \mathbb{E}_{s' \sim T(s,a), a' \sim \pi(s')} \left[Q_\pi(s', a') \right], \tag{6}$$

where $\mathcal{B}_\pi$ is the expected Bellman operator with respect to π. Thus, the critic is learned according to

$$\min_{Q_\theta} J_{\text{critic}}(Q_\theta) := \tfrac{1}{2} \mathbb{E}_{(s,a)\sim\beta} \left[(\mathcal{B}_\pi Q_\theta - Q_\theta)(s,a)^2 \right], \tag{7}$$

for some distribution β. Although the use of an arbitrary β suggests the critic may be learned off-policy, to achieve satisfactory performance, actor-critic algorithms generally rely on augmenting a replay buffer with *new* on-policy experience. Theoretical work has suggested that the 'ideal' β is, in fact, the on-policy distribution d^π (Sutton et al., 2000; Nachum et al., 2018b; Silver et al., 2014).

In this work, we focus on the off-policy setting directly. Specifically, we assume we are given a dataset $\mathcal{D} = \{(s_k, a_k, r_k, s'_k)\}_{k=1}^N$, where $r_k = r(s_k, a_k)$; $s'_k \sim T(s_k, a_k)$; and a_k has been sampled given s_k according to an unknown policy. We let $d^\mathcal{D}$ denote the unknown state-action distribution, and additionally assume access to a finite sample $\mathcal{U} = \{s_{0,k}\}_{k=1}^N$ such that $s_{0,k} \sim \mu_0$.

3 ALGAE Derived via State-Action Density Regularization

We begin by presenting an informal derivation of our method, motivated as a regularization of the dual max-return objective in (3). In Section 4 we will present our results more formally as a consequence of the Lagrangian of a linear programming formulation of the max-return objective.

3.1 A Regularized Off-Policy Max-Return Objective

Traditionally, optimizing the max-return objective requires access to on-policy samples from d^π. To introduce an off-policy distribution $d^\mathcal{D}$ to the objective, we add a regularizer:

$$\max_\pi J_{\text{D},f}(\pi) := \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] - \alpha D_f(d^\pi \| d^\mathcal{D}), \tag{8}$$

where $\alpha > 0$, f is some convex function, and D_f denotes the f-divergence:

$$D_f(d^\pi \| d^\mathcal{D}) = \mathbb{E}_{(s,a)\sim d^\mathcal{D}} \left[f\left(w_{\pi/\mathcal{D}}(s,a) \right) \right], \tag{9}$$

such that $w_{\pi/\mathcal{D}}(s,a) := \frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)}$. This form of regularization encourages *conservative* behavior, compelling the state-action occupancies of π to remain close to the off-policy distribution, although by appropriately choosing α and f, the strength of the regularization can be controlled. Later we show that many of our results also hold for *exploratory* regularization ($\alpha < 0$) and even for no regularization at all ($\alpha = 0$).

At first glance, the regularization seems to complicate things. Not only do we still require on-policy samples from d^π, but we also have to compute $D_f(d^\pi \| d^\mathcal{D})$, which in general can be difficult. To make this objective more approachable, we transform the f-divergence to its variational form by use of a dual function $x : S \times A \to \mathbb{R}$:

$$\begin{aligned}\max_\pi \min_{x: S\times A \to \mathbb{R}} J_{\text{D},f}(\pi, x) &:= \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] + \alpha \cdot \mathbb{E}_{(s,a)\sim d^\mathcal{D}}[f_*(x(s,a))] - \alpha \cdot \mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)] \\ &= \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a) - \alpha \cdot x(s,a)] + \alpha \cdot \mathbb{E}_{s,a\sim d^\mathcal{D}}[f_*(x(s,a))], \end{aligned} \tag{10}$$

where f_* is the convex (or Fenchel) conjugate of f. With the objective in (10), we are finally ready to eliminate the expectation over on-policy samples from d^π. To do so, we make a change of variables, inspired by the *DualDICE* trick (Nachum et al., 2019b). Define $\nu : S \times A \to \mathbb{R}$ as,

$$\nu(s,a) := -\alpha \cdot x(s,a) + \mathcal{B}_\pi \nu(s,a). \tag{11}$$

Applying this change of variables to 10 (after some telescoping, see Nachum et al. (2019b)) yields

$$\max_\pi \min_{\nu: S\times A \to \mathbb{R}} J_{\text{D},f}(\pi, \nu) := (1-\gamma) \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}} [\nu(s_0, a_0)] + \alpha \cdot \mathbb{E}_{(s,a)\sim d^\mathcal{D}}[f_*((\mathcal{B}_\pi \nu - \nu)(s,a)/\alpha)]. \tag{12}$$

The resulting objective is now completely off-policy, relying only on access to samples from the initial state distribution μ_0 and the off-policy dataset $d^\mathcal{D}$. Thus, we have our first theorem, providing an off-policy formulation of the max-return objective:

Theorem 1 (Primal ALGAE) *Under mild conditions on $d^\mathcal{D}$, α, f, the regularized max-return objective may be expressed as a max-min optimization:*

$$\begin{aligned} &\max_\pi \ \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] - \alpha D_f(d^\pi \| d^\mathcal{D}) \equiv \\ &\max_\pi \min_{\nu: S\times A \to \mathbb{R}} \ (1-\gamma) \mathbb{E}_{a_0 \sim \pi(s_0), s_0 \sim \mu_0}[\nu(s_0, a_0)] + \alpha \cdot \mathbb{E}_{(s,a)\sim d^\mathcal{D}}[f_*((\mathcal{B}_\pi \nu - \nu)(s,a)/\alpha)]. \end{aligned} \tag{13}$$

Remark (Fenchel ALGAE) The appearance of $\mathcal{B}_\pi$ inside f_* in the second term of (13) presents a challenge in practice, since $\mathcal{B}_\pi$ involves an expectation over the transition function T, whereas

one typically only has access to a single empirical sample from T. To handle this issue, one can apply the dual embedding technique (Dai et al., 2016, 2017). Specifically, the dual representation of f_*, $f_*((\mathcal{B}_\pi\nu - \nu)(s,a)/\alpha) = \max_\zeta \frac{1}{\alpha}(\mathcal{B}_\pi\nu - \nu)(s,a)\cdot\zeta - f(\zeta)$, can be substituted into (13), to achieve the $\max$-$\min$-$\max$ optimization:

$$\max_\pi \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] - \alpha D_f(d^\pi \| d^\mathcal{D}) \equiv \max_\pi \min_{\nu: S\times A\to\mathbb{R}} \max_{\zeta: S\times A\to\mathbb{R}} (1-\gamma)\mathbb{E}_{\substack{a_0\sim\pi(s_0)\\ s_0\sim\mu_0}}[\nu(s_0,a_0)] + \mathbb{E}_{(s,a)\sim d^\mathcal{D}, s'\sim T(s,a), a'\sim\pi(s')}[(\gamma\nu(s',a') - \nu(s,a))\cdot\zeta(s,a) - \alpha\cdot f(\zeta(s,a))]. \quad (14)$$

As we will see in Section 4, under mild conditions, strong duality holds in the inner $\min$-$\max$ of (14), hence one can switch the $\min_\nu$ and $\max_\zeta$ to reduce to a more convenient $\max$-$\max$-$\min$ form.

Remark (Donsker-Varadhan regularization) We note that the use of an f-divergence regularization can be replaced with other regularizations that possess a variational form with a linear expectation $\mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)]$. One particularly interesting intantiation of this is given by the Donsker-Varadhan representation of the KL-divergence (Donsker and Varadhan, 1983):

$$-D_{\mathrm{KL}}(d^\pi || d^\mathcal{D}) = \min_{xS\times A\to\mathbb{R}} \log \mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[\exp\{x(s,a)\}\right] - \mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)]. \quad (15)$$

This form of regularization can be applied to analogous derivations to yield a variant of ALGAE with a log-average-exp term.

3.2 On-Policy Policy Gradient using Off-Policy Data

The equivalence between the objective in (12) and the on-policy max-return objective can be highlighted by considering the gradient of this objective with respect to π. First, consider the optimal $x^*_\pi := \mathrm{argmin}_x J_{\mathrm{D},f}(\pi, x)$ for (10). By taking the gradient of $J_{\mathrm{D},f}$ with respect to x and setting this to 0, one finds that x^*_π satisfies

$$f'_*(x^*_\pi(s,a)) = w_{\pi/\mathcal{D}}(s,a). \quad (16)$$

Accordingly, for any π, the optimal $\nu^*_\pi := \mathrm{argmin}_\nu J_{\mathrm{D},f}(\pi,\nu)$ for (12) satisfies

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

Thus, we may express the gradient of $J_{\mathrm{D},f}(\pi,\nu^*_\pi)$ with respect to π as

$$\begin{aligned}\frac{\partial}{\partial\pi}J_{\mathrm{D},f}(\pi,\nu^*_\pi) &= (1-\gamma)\frac{\partial}{\partial\pi}\mathbb{E}_{\substack{a_0\sim\pi(s_0)\\ s_0\sim\mu_0}}[\nu^*_\pi(s_0,a_0)] + \mathbb{E}_{(s,a)\sim d^\mathcal{D}}\left[w_{\pi/\mathcal{D}}(s,a)\frac{\partial}{\partial\pi}(\mathcal{B}_\pi\nu^*_\pi - \nu^*_\pi)(s,a)\right]\\ &= (1-\gamma)\frac{\partial}{\partial\pi}\mathbb{E}_{\substack{a_0\sim\pi(s_0)\\ s_0\sim\mu_0}}[\nu^*_\pi(s_0,a_0)] + \gamma\cdot\mathbb{E}_{\substack{(s,a)\sim d^\pi,\\ s'\sim T(s,a)}}\left[\frac{\partial}{\partial\pi}\mathbb{E}_{a'\sim\pi(s')}[\nu^*_\pi(s',a')]\right]\\ &= \mathbb{E}_{(s,a)\sim d^\pi}\left[\nu^*_\pi(s,a)\nabla\log\pi(a|s)\right],\end{aligned}$$

where the second equality comes from Danskin's theorem (Bertsekas, 1999). This means that, if the dual function ν is sufficiently optimized, the gradient of the off-policy objective $J_{\mathrm{D},f}(\nu,\pi)$ is *exactly* the on-policy policy gradient, with Q-value function given by ν^*_π.

To characterize this Q-value function, note that from (11), ν^*_π is a Q-value function with respect to augmented reward $\tilde{r}(s,a) := r(s,a) - \alpha\cdot x^*_\pi(s,a)$. Recalling the expression for x^*_π in (16) and the fact that, for any convex f, the derivatives f' and f'_* are inverses of each other, we have, $\tilde{r}(s,a) = r(s,a) - \alpha\cdot f'(w_{\pi/\mathcal{D}}(s,a))$. This derivation leads to our second theorem:

Theorem 2 *If the dual function ν is sufficiently optimized, the gradient of the off-policy objective $J_{\mathrm{D},f}(\pi,\nu)$ with respect to π is exactly the (regularized) on-policy policy gradient:*

$$\frac{\partial}{\partial\pi}\min_\nu J_{\mathrm{D},f}(\pi,\nu) = \mathbb{E}_{(s,a)\sim d^\pi}\left[\tilde{Q}^\pi(s,a)\nabla\log\pi(a|s)\right], \quad (18)$$

where, $\tilde{Q}^\pi(s,a)$ is the Q-value function of π with respect to rewards $\tilde{r}(s,a) := r(s,a) - \alpha\cdot f'(w_{\pi/\mathcal{D}}(s,a))$.

Remark We note that Theorem 2 also holds when using the more sophisticated objective in (14), since the optimal ζ^*_π is equal to $w_{\pi/\mathcal{D}}$, regardless of π.

3.3 Connection to Actor-Critic

The relationship between the proposed off-policy objective and the classical policy gradient becomes more profound when we consider the form of the objective under specific choices of convex function

f. If we take $f(x) = \frac{1}{2}x^2$, then $f_*(x) = \frac{1}{2}x^2$ and the proposed objective is reminiscent of actor-critic:

$$\max_{\pi} \min_{\nu: S\times A\to\mathbb{R}} J_{\mathrm{D},f}(\pi,\nu) := (1-\gamma)\mathbb{E}_{s_0\sim\mu_0,a_0\sim\pi(s_0)}[\nu(s_0,a_0)] + \frac{1}{2\alpha}\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[((\mathcal{B}_\pi\nu-\nu)(s,a))^2].$$

The second term alone is an instantiation of the off-policy critic objective in actor-critic. However, in actor-critic, the use of an off-policy objective for the critic is difficult to theoretically motivate. Moreover, in practice, critic and actor learning can both suffer from the mismatch between the off-policy distribution $d^{\mathcal{D}}$ and the on-policy d^π. By contrast, our derivations show that the introduction of the first term to the objective transforms the off-policy actor-critic algorithm to an on-policy actor-critic, without any explicit use of importance weights. Moreover, while standard actor-critic has two separate objectives for value and policy, our proposed objective is a single, unified objective. Both the policy and value functions are trained with respect to the same off-policy objective.

4 A Lagrangian View of ALGAE

We now show how ALGAE can be alternatively derived from the Lagrangian of a linear programming (LP) formulation of the Q_π-function. Please refer to Appendix A for details. We begin by formalizing an LP characterization of the Q_π-function, with its dual form:

Theorem 3 *Given a policy π, the average return of π may be expressed in primal and dual forms as*

$$\min_{\nu: S\times A\to\mathbb{R}} J_{\mathrm{P}}(\pi,\nu) := (1-\gamma)\,\mathbb{E}_{\mu_0\pi}\left[\nu\left(s_0,a_0\right)\right] \quad \text{s.t.} \quad \nu\left(s,a\right) \geq \mathcal{B}_\pi\nu(s,a), \quad \forall\left(s,a\right)\in S\times A, \tag{19}$$

and,

$$\max_{\rho: S\times A\to\mathbb{R}_+} J_{\mathrm{D}}(\pi,\rho) := \mathbb{E}_\rho\left[r\left(s,a\right)\right] \quad \text{s.t.} \quad \rho\left(s,a\right) = \mathcal{B}_\pi^T\rho(s,a), \quad \forall\left(s,a\right)\in S\times A, \tag{20}$$

respectively, where $\mathcal{B}_\pi^T\rho(s,a) := (1-\gamma)\mu_0(s)\pi(a|s) + \gamma\sum_{\tilde{s},\tilde{a}}\pi(\tilde{a}|\tilde{s})T(s|\tilde{s},\tilde{a})\rho(\tilde{s},\tilde{a})$ is the transpose Bellman operator. *The optimal primal ν_π^* is Q_π and the optimal dual ρ_π^* is d^π.*

Consider the Lagrangian of J_{P}, which would typically be expressed with a sum (or integral) of constraints weighted by ρ. Under the assumption that $d^{\mathcal{D}}$ covers the support of optimal dual variables d^π, we can reparametrize the dual variable as $\zeta\left(s,a\right) = \frac{\rho(s,a)}{d^{\mathcal{D}}(s,a)}$ to express the Lagrangian as,

$$\min_{\nu: S\times A\to\mathbb{R}} \max_{\zeta: S\times A\to\mathbb{R}_+} (1-\gamma)\,\mathbb{E}_{s_0\sim\mu_0,a_0\sim\pi(s_0)}\left[\nu\left(s_0,a_0\right)\right] + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\zeta\left(s,a\right)\left(\mathcal{B}_\pi\nu-\nu\right)(s,a)\right]. \tag{21}$$

The optimal ζ_π^* of this Lagrangian is $w_{\pi/\mathcal{D}}$. In practice, the linear structure in (21) can induce numerical instability. Therefore, inspired by the augmented Lagrangian method, we introduce regularization. By adding a special regularizer $\alpha\cdot\mathbb{E}_{d^{\mathcal{D}}}\left[f\left(\zeta\left(s,a\right)\right)\right]$ using a convex f, we obtain

$$\min_{\nu}\max_{\zeta} L\left(\nu,\zeta;\pi\right) := (1-\gamma)\,\mathbb{E}_{s_0\sim\mu_0,a_0\sim\pi(s_0)}\left[\nu\left(s_0,a_0\right)\right] + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\zeta\left(s,a\right)\left(\mathcal{B}_\pi\nu-\nu\right)(s,a)\right] - \alpha\cdot\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[f\left(\zeta\left(s,a\right)\right)\right]. \tag{22}$$

We now characterize the optimizers ν_π^* and ζ_π^* and the optimum value $L(\nu_\pi^*,\zeta_\pi^*;\pi)$ of this objective.

Theorem 4 *If $w_{\pi/\mathcal{D}}$ is bounded and the convex (Fenchel) conjugate f_* of f is closed and strongly convex, then the solution to* (22) *is given by,*

$$\nu_\pi^*\left(s,a\right) = -\alpha f'\left(w_{\pi/\mathcal{D}}\left(s,a\right)\right) + \mathcal{B}_\pi\nu_\pi^*(s,a), \text{ and } \zeta_\pi^*\left(s,a\right) = w_{\pi/\mathcal{D}}\left(s,a\right).$$

The optimal value is $L\left(\nu_\pi^*,\zeta_\pi^*;\pi\right) = \mathbb{E}_{d^\pi}[r(s,a)] - \alpha D_f(d^\pi\|d^{\mathcal{D}})$.

Thus, we have recovered the Fenchel ALGAE objective for π in the remark following Theorem 1 (Equation 14). To recover the result of Theorem 1 exactly (Primal ALGAE, Equation 13), one simply undoes the application of Fenchel duality to recover $f_*((\mathcal{B}_\pi\nu-\nu)(s,a)/\alpha) = \max_\zeta\ \frac{1}{\alpha}(\mathcal{B}_\pi\nu-\nu)(s,a)\cdot\zeta - f(\zeta)$.

The derivation of this same result from the LP perspective allows us to exploit strong duality. Specifically, under the assumption that $w_{\pi/\mathcal{D}}(s,a)$ and $r\left(s,a\right)$ are bounded, then (ν_π^*,ζ_π^*) does not change if we optimize $L\left(\nu,\zeta;\pi\right)$ over a bounded space $\mathcal{H}\times\mathcal{F}$, provided $(\nu_\pi^*,\zeta_\pi^*)\in\mathcal{H}\times\mathcal{F}$. In this case strong duality holds by Proposition 2.1 in Ekeland and Temam (1999), and we obtain

$$\min_{\nu\in\mathcal{H}}\max_{\zeta\in\mathcal{F}} L\left(\nu,\zeta;\pi\right) = \max_{\zeta\in\mathcal{F}}\min_{\nu\in\mathcal{H}} L\left(\nu,\zeta;\pi\right).$$

This implies that, for computational efficiency, we can optimize the policy via

$$\max_{\pi\in\mathcal{P}}\ell\left(\pi\right) := \max_{\zeta\in\mathcal{F}}\min_{\in\mathcal{H}}\,(1-\gamma)\,\mathbb{E}_{\mu_0\pi}\left[\nu\left(s,a\right)\right] + \mathbb{E}_{d^{\mathcal{D}}}\left[\zeta\left(s,a\right)\left(\mathcal{B}_\pi\nu-\nu\right)(s,a)\right] - \alpha\mathbb{E}_{d^{\mathcal{D}}}\left[f\left(\zeta\left(s,a\right)\right)\right]. \tag{23}$$

Remark (extensions to $\gamma = 1$ or $\alpha \leq 0$): Although ALGAE is originally derived with $\gamma \in (0, 1)$ and $\alpha > 0$ in Section 3, the Lagrangian view of the LP formulation of Q_π can be used to generalize the algorithm to $\gamma = 1$ and $\alpha = 0$. In particular, for $\alpha = 0$, one can directly use the original Lagrangian for the LP. For the case $\gamma = 1$, the problem reduces to the Lagrangian of the LP for an undiscounted Q_π-function, as proved in Appendix A. Lastly, one may extend these results to $\alpha < 0$ (exploratory regularization) by simply changing the regularization from $-\alpha f(\zeta)$ to $\alpha f(\zeta)$ and changing the optimization over ζ from a $\max$ to a $\min$.

5 Related Work

Adding regularization to MDPs (Neu et al., 2017) has been investigated for many different purposes in the literature, including exploration (de Farias and Van Roy, 2000; Haarnoja et al., 2017, 2018a), smoothing (Dai et al., 2017), avoiding premature convergence (Nachum et al., 2017a), ensuring tractability (Todorov, 2006), and mitigating observation noise (Rubin et al., 2012; Fox et al., 2016). Plenty of classical RL algorithms have also been extended to consider such entropy-regularized MDPs, including SARSA (Asadi and Littman, 2017), Q-learning (Fox et al., 2016; Haarnoja et al., 2017), and actor-critic (Haarnoja et al., 2018a; Nachum et al., 2017a). We note that the regularization employed by ALGAE as a divergence over state-action densities is markedly different from these previous works, which mostly regularize the action distributions of a policy conditioned on state. Moreover, the various off-policy policy improvement algorithms that have been recently introduced for entropy-regularized MDPs (Nachum et al., 2017a, 2018a; Dai et al., 2017; Haarnoja et al., 2018a) ignore the bias induced by the distribution shift between the collected off-policy data and the state-action occupancy distribution of the current policy. Because of this mismatch, these algorithms still require a large amount of online interaction to achieve success in practice. By contrast, the proposed ALGAE approach relies on the variational form of an f-divergence regularization, which allows us to automatically compensate for this distribution shift.

Algorithmically, our proposed method follows a Lagrangian primal-dual view of the LP characterization of the Q-function, which leads to a saddle-point problem. Several recent works (Chen and Wang, 2016; Wang, 2017; Dai et al., 2017, 2018; Chen et al., 2018; Lee and He, 2018) have also considered saddle-point formulations for policy improvement, but these formulations are derived from fundamentally different perspectives. In particular, Dai et al. (2018) exploits a saddle-point formulation for the path-consistency conditions, while Chen and Wang (2016); Wang (2017); Dai et al. (2017); Chen et al. (2018) consider the (augmented) Lagrangian of the LP characterization of Bellman optimality for the optimal V-function, and Lee and He (2018) introduces auxiliary Q into the vanilla LP for the optimal V-function. By contrast, ALGAE is derived by expressing the novel LP characterization of the Q-function in an off-policy form, which allows us to achieve an *unbiased* policy gradient formulation from behavior-agnostic and off-policy data.

6 Experiments

We evaluate our algorithm in the online regime – environment steps are alternated with policy updates – and the offline regime – a fixed dataset of 5k transitions collected by an optimal policy is used. We compare the performance of ALGAE on the Pendulum compared to SAC in these settings (Figure 1). We find that ALGAE can achieve good performance on this task.

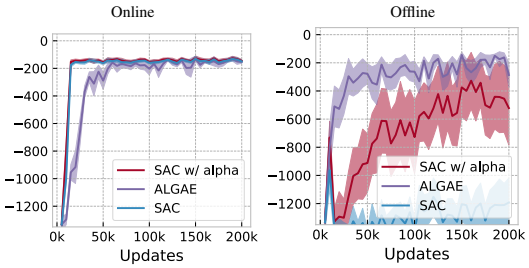

Figure 1: ALGAE on Pendulum.

7 Conclusion

We have introduced an ALgorithm for policy Gradient from Arbitrary Experience (ALGAE) to address the *behavior-agnostic off-policy* policy improvement task in reinforcement learning. Based on a novel linear programming characterization of the Q-function, we derived the new approach from a Lagrangian saddle-point formulation. The resulting algorithm, ALGAE, automatically compensates for the distribution shift in collected off-policy data, and achieves an estimate of the on-policy policy gradient using this off-policy data.

References

Kavosh Asadi and Michael L. Littman. An alternative softmax operator for reinforcement learning. In *ICML*, pages 243–252, 2017.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, second edition, 1999.

Yichen Chen and Mengdi Wang. Stochastic primal-dual methods and sample complexity of reinforcement learning. *arXiv preprint arXiv:1612.02516*, 2016.

Yichen Chen, Lihong Li, and Mengdi Wang. Scalable bilinear π learning using state and action features. *arXiv preprint arXiv:1804.10328*, 2018.

Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual embeddings. *CoRR*, abs/1607.04579, 2016.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. *CoRR*, abs/1712.10285, 2017.

Bo Dai, Albert Shaw, Niao He, Lihong Li, and Le Song. Boosting the actor with dual critic. *ICLR*, 2018. arXiv:1712.10282.

Daniela Pucci de Farias and Benjamin Van Roy. On the existence of fixed points for approximate value iteration and temporal-difference learning. *Journal of Optimization Theory and Applications*, 105(3):589–608, 2000.

Thomas Degris, Martha White, and Richard S Sutton. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*, volume 28. Siam, 1999.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In *UAI*, 2016.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. *arXiv preprint arXiv:1702.08165*, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018a.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018b.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

Donghwan Lee and Niao He. Stochastic primal-dual q-learning. *arXiv preprint arXiv:1810.08298*, 2018.

Tyler Lu, Dale Schuurmans, and Craig Boutilier. Non-delusional q-learning and value-iteration. In *Advances in Neural Information Processing Systems*, pages 9949–9959, 2018.

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.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. *arXiv preprint arXiv:1702.08892*, 2017a.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Trust-pcl: An off-policy trust region method for continuous control. *arXiv preprint arXiv:1707.01891*, 2017b.

Ofir Nachum, Yinlam Chow, and Mohammad Ghavamzadeh. Path consistency learning in tsallis entropy regularized mdps. *arXiv preprint arXiv:1802.03501*, 2018a.

Ofir Nachum, Mohammad Norouzi, George Tucker, and Dale Schuurmans. Smoothed action value functions for learning gaussian policies. *arXiv preprint arXiv:1803.02348*, 2018b.

Ofir Nachum, Michael Ahn, Hugo Ponte, Shixiang Gu, and Vikash Kumar. Multi-agent manipulation via locomotion using hierarchical sim2real. *arXiv preprint arXiv:1908.05224*, 2019a.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Efficient estimation of off-policy stationary distribution corrections. 2019b.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

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

Jonathan Rubin, Ohad Shamir, and Naftali Tishby. Trading value and information in MDPs. *Decision Making with Imperfect Decision Makers*, pages 57–74, 2012.

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

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. 2014.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

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.

Emanuel Todorov. Linearly-solvable Markov decision problems. In *NIPS*, pages 1369–1376, 2006.

Mengdi Wang. Randomized Linear Programming Solves the Discounted Markov Decision Problem In Nearly-Linear Running Time. *ArXiv e-prints*, 2017.

Tao Wang, Daniel Lizotte, Michael Bowling, and Dale Schuurmans. Dual representations for dynamic programming. 2008.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016.

Appendix

A Proof Details

We follow the notations in main text. For convenience, we will use $\sum$ and $\int$ interchangeably.

Theorem 3 *Given a policy π, the average return of π may be expressed in primal and dual forms as,*

$$\min_{\nu:S\times A\to\mathbb{R}} J_{\mathrm{P}}(\pi,\nu) := (1-\gamma)\,\mathbb{E}_{\mu_0\pi}\left[\nu\left(s_0,a_0\right)\right] \quad \text{s.t.} \quad \nu\left(s,a\right) \geq \mathcal{B}_\pi\nu(s,a), \ \forall\left(s,a\right)\in S\times A, \tag{24}$$

and,

$$\max_{\rho:S\times A\to\mathbb{R}_+} J_{\mathrm{D}}(\pi,\rho) := \mathbb{E}_\rho\left[r\left(s,a\right)\right] \quad \text{s.t.} \quad \rho\left(s,a\right) = \mathcal{B}_\pi^T\rho(s,a), \ \forall\left(s,a\right)\in S\times A, \tag{25}$$

respectively, where $\mathcal{B}_\pi^T\rho(s,a) := (1-\gamma)\mu_0(s)\pi(a|s) + \gamma\sum_{\tilde{s},\tilde{a}}\pi(\tilde{a}|\tilde{s})T(s|\tilde{s},\tilde{a})\rho(\tilde{s},\tilde{a})$ is the transpose Bellman operator. *The optimal primal ν_π^* is Q_π and the optimal dual ρ_π^* is d^π.*

Proof Recall we have $\mathcal{B}_\pi$ is monotonic, *i.e.*, if $\nu \geq \mathcal{B}_\pi\nu \Rightarrow \mathcal{B}_\pi\nu \geq \left(\mathcal{B}_\pi\right)^2\nu$ and $Q_\pi = \left(\mathcal{B}_\pi\right)^\infty\nu$ for arbitrary ν, we have for $\forall\nu$ feasible, $\nu \geq \left(\mathcal{B}_\pi\right)\nu \geq \left(\mathcal{B}_\pi\right)^2\nu \geq \left(\mathcal{B}_\pi\right)^3\nu \geq \ldots \geq \left(\mathcal{B}_\pi\right)^\infty\nu = Q_\pi$, which achieves the first claim.

The duality of the linear program (19) can be easily obtained as

$$\begin{aligned} \max_{\rho:S\times A\to\mathbb{R}_+} \quad & \mathbb{E}_\rho\left[R\left(s,a\right)\right], \\ \text{s.t.} \quad & \rho\left(s',a'\right) - \gamma\sum_{s,a}\pi\left(a'|s'\right)T\left(s'|s,a\right)\rho\left(s,a\right) = (1-\gamma)\,\mu_0\left(s'\right)\pi\left(a'|s'\right), \\ & \forall\left(s',a'\right)\in S\times A. \end{aligned} \tag{26}$$

Notice that there are $|S|\times|A|$ equality constraints, the only solution to (26) can be obtained by solving linear system formed by the constraints, leading to

$$\rho^* = (1-\gamma)\left(I-\gamma\left(P_\pi\right)^\top\right)^{-1}(\mu\pi), \tag{27}$$

where $P_\pi\left(s',a'|s,a\right) = \pi\left(a'|s'\right)T\left(s'|s,a\right)$. Recall that when $\gamma\in(0,1)$, we can rewrite $\left(I-\gamma\left(P_\pi\right)^\top\right)^{-1} = \sum_{t=0}^\infty\gamma^t\left(P_\pi\right)^t$, we have $\rho^* = (1-\gamma)\sum_{t=0}^\infty\gamma^t\left(P_\pi\right)^t(\mu\pi) = d^\pi$, which is the stationary occupancy of π. ■

Remark ($\gamma = 1$ undiscounted MDP): for $\gamma = 1$, (19) is no longer valid. However, the dual problem (26) may be modified by simply adding a constraint $\sum_{s,a}\rho\left(s,a\right) = 1$ to achieve the same result. Therefore, one can derive the analogous Lagrangian for $\gamma = 1$ as,

$$\begin{aligned} \max_{\zeta\in Z}\min_{\nu:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \quad & L\left(\zeta,\nu;\pi\right) := \mathbb{E}_{d^{\mathcal{D}},\pi}\left[\zeta\left(s,a\right)\nu\left(s',a'\right)\right] + \mathbb{E}_{d^{\mathcal{D}}}\left[\zeta\left(s,a\right)\left(\nu\left(s,a\right) - r\left(s,a\right)\right)\right] \\ \text{s.t.} \quad & \zeta\in Z := \left\{\zeta\geq 0, \mathbb{E}_{d^{\mathcal{D}}}\left[\zeta\left(s,a\right)\right] = 1\right\}. \end{aligned} \tag{28}$$

By the KKT condition, the optimal value of the Lagrangian is $\mathbb{E}_{d^\pi}\left[r\left(s,a\right)\right]$, which is the policy value of π.

Therefore, for both $\gamma\in(0,1)$ and $\gamma = 1$, we can improve the policy by maximizing π in the (regularized) Lagrangian.

Theorem 4 *If $w_{\pi/\mathcal{D}}$ is bounded and the convex (Fenchel) conjugate f_* of f is closed and strongly convex, then the solution to* (22) *is given by,*

$$\nu_\pi^*\left(s,a\right) = -\alpha f'\left(w_{\pi/\mathcal{D}}\left(s,a\right)\right) + \mathcal{B}_\pi\nu_\pi^*(s,a), \quad \textit{and} \quad \zeta_\pi^*\left(s,a\right) = w_{\pi/\mathcal{D}}\left(s,a\right).$$

The optimal value is $L\left(\nu_\pi^*,\zeta_\pi^*;\pi\right) = \mathbb{E}_{d^\pi}[r(s,a)] - \alpha D_f(d^\pi\|d^{\mathcal{D}})$.

Proof

By Fenchel duality, we have

$$\max_{\zeta:\mathcal{S}\times\mathcal{A}\to\mathbb{R}_+}\ \mathbb{E}_{d^{\mathcal{D}}}\left[\zeta\left(s,a\right)\left(\mathcal{B}_\pi\nu-\nu\right)\left(s,a\right)\right] - \alpha\mathbb{E}_{d^{\mathcal{D}}}\left[f\left(\zeta\left(s,a\right)\right)\right] = \alpha\mathbb{E}_{d^{\mathcal{D}}}\left[f_*\left(\frac{1}{\alpha}\left(\mathcal{B}_\pi\nu-\nu\right)\left(s,a\right)\right)\right],$$

where $f_*(\cdot)$ denotes the conjugate function of $f(\cdot)$ over $\mathbb{R}_+$. Plugging this into the optimization, we have,

$$L(\nu, \zeta_\pi^*; \pi) = \min_{\nu: \mathcal{S}\times\mathcal{A}\to\mathbb{R}} (1-\gamma)\,\mathbb{E}_{\mu_0\pi}\left[\nu(s_0, a_0)\right] + \alpha\mathbb{E}_{d^\mathcal{D}}\left[f_*\left(\frac{1}{\alpha}\left(\mathcal{B}_\pi - I\right)\circ\nu(s,a)\right)\right]. \quad (29)$$

To investigate the optimality, we apply the change-of-variable, *i.e.*, $x(s,a) = \frac{1}{\alpha}\left(\mathcal{B}_\pi - I\right)\circ\nu(s,a)$. Let $\beta_t(s) = P\left(s = s_t | s_0 \sim \mu_0, \{a_i\}_{i=0}^t \sim \pi\right)$. Then, the first expectation in (29) may be expressed as,

$$\begin{aligned}
&(1-\gamma)\mathbb{E}_{\mu_0\pi}[\nu(s_0,a_0)] \\
&= (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathbb{E}_{s\sim\beta_t, a\sim\pi(s)}[\nu(s,a)] - (1-\gamma)\sum_{t=0}^\infty \gamma^{t+1}\mathbb{E}_{s'\sim\beta_{t+1}, a'\sim\pi(s')}[\nu(s',a')] \\
&= \mathbb{E}_{(s,a)\sim d^\pi}[\nu(s,a) - \gamma\mathbb{E}_{s'\sim T(s,a), a'\sim\pi(s')}[\nu(s',a')]] \\
&= \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] + \mathbb{E}_{(s,a)\sim d^\pi}[\nu(s,a) - r(s,a) - \gamma\mathbb{E}_{s'\sim T(s,a), a'\sim\pi(s')}[\nu(s',a')]] \\
&= \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] - \alpha\mathbb{E}_{(s,a)\sim d^\pi}[x(s,a)].
\end{aligned}$$

Let $\mathcal{C}$ denote the set of functions x in the image of $(\mathcal{B}_\pi - I)\circ\nu$ for $\nu: S\times A\to\mathbb{R}$. Therefore, the change of variables yields the following re-formulation of L:

$$\begin{aligned}
L(\nu_\pi^*, \zeta_\pi^*; \pi) &= \min_{x\in\mathcal{C}} \mathbb{E}_{(s,a)\sim d^\pi}\left[r(s,a)\right] - \alpha\mathbb{E}_{(s,a)\sim d^\pi}\left[x(s,a)\right] + \alpha\mathbb{E}_{d^\mathcal{D}}\left[f_*\left(x(s,a)\right)\right] \\
&= \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] - \alpha\left(\max_{x\in\mathcal{C}} \mathbb{E}_{d^\pi}\left[x(s,a)\right] - \mathbb{E}_{d^\mathcal{D}}\left[f_*\left(x(x,a)\right)\right]\right)
\end{aligned}$$

Note that, ignoring the restriction of x to $\mathcal{C}$, the optimal x_π^* satisfies $f_*'(x(s,a)) = w_{\pi/\mathcal{D}}(s,a)$. Under the assumption that $f_*(\cdot)$ is strongly convex, then $\left[(f_*)'\right]^{-1}(\cdot)$ exists, and it is given by $f'(\cdot)$. Thus, we have $x_\pi^*(s,a) = f'(w_{\pi/\mathcal{D}}(s,a))$ for all s, a. Due to the assumption that $w_{\pi/\mathcal{D}}$ is bounded, we have that x_π^* is bounded and thus $x_\pi^* \in \mathcal{C}$. Therefore, by definition of the f-divergence, we have,

$$L(\nu_\pi^*, \zeta_\pi^*; \pi) = \mathbb{E}_{(s,a)\sim d^\pi}[r(s,a)] - \alpha D_f(d^\pi || d^\mathcal{D}), \quad (30)$$

as desired.

To characterize ν_π^*, we note,

$$x^*(s,a) = f'(w_{\pi/\mathcal{D}}(s,a)) \Rightarrow \nu_\pi^*(s,a) = \mathcal{B}_\pi\nu_\pi^*(s,a) - \alpha f'(w_{\pi/\mathcal{D}}(s,a)). \quad (31)$$

On the other hand, for the optimal dual $\zeta_\pi^*(s,a)$, we have

$$\zeta_\pi^*(s,a) = \operatorname*{argmin}_\zeta \zeta\cdot x_\pi^*(s,a) - f(\zeta) = f_*'(x_\pi^*(s,a)) = w_{\pi/\mathcal{D}}(s,a)$$

where the second equality comes from the fact that $f'(\zeta_\pi^*(s,a)) = x_\pi^*(s,a) \Rightarrow \zeta_\pi^*(s,a) = f_*'(x_\pi^*(s,a))$.

■

B Experiment Details

For ALGAE we used a primal version with log-average-exp representation of KL. We perform mini-batch training with Adam with batch size 256. For both SAC and ALGAE, we used learning rate 10^{-3} and 10^{-4} for the policy networks and the critic/ν networks respectively. For every policy update we performed 2 critic/ν updates (as in TD3). We set SAC temperature to 1, while ALGAE α to 1. We performed deterministic evaluation of the algorithm by taking mode actions.