REGULARIZED INVERSE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Inverse Reinforcement Learning (IRL) aims to facilitate a learner's ability to imitate expert behavior by acquiring reward functions that explain the expert's decisions. *Regularized IRL* applies convex regularizers to the learner's policy in order to avoid degenerate solutions, i.e., whereby any constant reward rationalizes the expert behavior. We propose analytical solutions, and practical methods to obtain them, for regularized IRL. Current methods are restricted to the maximum-entropy IRL framework, limiting them to Shannon-entropy regularizers, as well as proposing functional-form solutions that are generally intractable. We present theoretical backing for our proposed IRL method's applicability for both discrete and continuous controls, empirically validating our performance on a variety of tasks.

1 INTRODUCTION

Reinforcement learning (RL) has been successfully applied to many challenging domains including games (Mnih et al., 2015; 2016) and robot control (Schulman et al., 2015; Fujimoto et al., 2018; Haarnoja et al., 2018). Advanced RL methods often employ policy regularization motivated by, e.g., boosting exploration (Haarnoja et al., 2018) or safe policy improvement (Schulman et al., 2015). While Shannon entropy is often used as a policy regularizer (Ziebart et al., 2008), Geist et al. (2019) recently proposed a theoretical foundation of *regularized Markov decision processes* (MDPs) – a framework that uses arbitrary convex functions as policy regularizers. Here, one crucial advantage is that an optimal policy can be shown to *uniquely* exists, whereas multiple optimal policies may exist in the absence of policy regularization (and depending on the given reward structure).

Meanwhile, since RL requires a given or known reward function (which can often involve non-trivial reward engineering), Inverse Reinforce Learning (IRL) (Russell, 1998; Ng et al., 2000) – the problem of acquiring a reward function that promotes expert-like behavior – is more generally adopted in practical scenarios like robotic manipulation (Finn et al., 2016b), autonomous driving (Sharifzadeh et al., 2016; Wu et al., 2020) and clinical motion analysis (Li et al., 2018). In these scenarios, defining a reward function beforehand is particularly challenging and IRL is simply more pragmatic. However, complications with IRL in unregularized MDPs relate to the issue of degeneracy and early work has shown that degenerate solutions, where any constant function can be a solution of the expert's behavior, can lead to suboptimal learned policies (Ng et al., 2000).

Fortunately, Geist et al. (2019) show that IRL in regularized MDPs – *regularized IRL* – do not contain such degenerate solutions due to the uniqueness of the optimal policy for regularized MDPs. Despite this, no analytical solutions of regularized IRL – other than maximum-Shannon-entropy IRL (MaxEntIRL) (Ziebart et al., 2008; Ziebart, 2010; Ho & Ermon, 2016; Finn et al., 2016a; Fu et al., 2018) – have been proposed.

Solutions of a functional form were proposed in Geist et al. (2019), however they are generally intractable since a closed-form relation between the policy and optimal value function is needed to derive their solutions. This is not applicable with arbitrary policy regularization. Furthermore, practical algorithms for solving regularized IRL problems have not yet been proposed.

We summarize our contribution as follows: unlike the solutions in Geist et al. (2019), we propose tractable solutions for regularized IRL problems (Section 3) that require only knowledge of the analytical form of policy regularization. We show that our solutions are valid in continuous control problems including solutions for Tsallis entropy regularization with multi-variate Gaussian policies (Section 3.2). We devise Regularized Adversarial Inverse Reinforcement Learning (RAIRL), a

practical sample-based method for policy imitation and reward learning in regularized MDPs, which generalizes adversarial IRL (AIRL, Fu et al. (2018)) (Section 4). Finally, we empirically validate our RAIRL method on both discrete and continuous control tasks, evaluating both episodic scores and divergence minimization perspective Ghasemipour et al. (2019); Dadashi et al. (2020) (Section 5).

2 PRELIMINARIES

Notation For finite sets X and Y, Y^X is a set of functions from X to Y. Δ^X (Δ^X_Y) is a set of (conditional) probabilities over X (conditioned on Y). Especially for the conditional probabilities $p \in \Delta^X_Y$, we say $p(\cdot|y) \in \Delta^X$ for $y \in Y$. $\mathbb{R}$ is the set of real numbers. For functions $f_1, f_2 \in \mathbb{R}^X$, the inner product between f_1 and f_2 on X is defined as $\langle f_1, f_2 \rangle_X := \sum_{x \in X} f_1(x)\, f_2(x)$.

Regularized Markov Decision Processes and Reinforcement Learning We consider sequential decision making problems where an agent sequentially chooses its actions after observing the state of the environment, and the environment in turn emits a reward and state transition. Such an interaction between the agent and the environment is modeled as an infinite-horizon MDP, $\mathcal{M}^r := \langle \mathcal{S}, \mathcal{A}, P_0, P, r, \gamma \rangle$ and the agent's policy $\pi \in \Delta^{\mathcal{A}}_{\mathcal{S}}$. The terms within the MDP are defined as follows: $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $P_0 \in \Delta^{\mathcal{S}}$ is an initial state distribution, $P \in \Delta^{\mathcal{S}}_{\mathcal{S}\times\mathcal{A}}$ is a state transition probability, $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is a reward function, and $\gamma \in [0, 1)$ is the discount factor. We also define an MDP without reward as $\mathcal{M}^- := \langle \mathcal{S}, \mathcal{A}, P_0, P, \gamma \rangle$. The normalized state-action visitation distribution, $d_\pi \in \Delta^{\mathcal{S}\times\mathcal{A}}$, associated with π is defined as the expected discounted state-action visitation of π, i.e., $d_\pi(s, a) := (1-\gamma) \cdot \mathbb{E}_\pi[\sum_{i=0}^{\infty} \gamma^i \mathbb{I}\{s_i = s, a_i = a\}]$, where the subscript π on $\mathbb{E}$ means that a trajectory $(s_0, a_0, s_1, a_1, ...)$ is randomly generated from $\mathcal{M}^-$ and π, and $\mathbb{I}\{\cdot\}$ is an indicator function. Note that d_π satisfies the transposed Bellman recurrence (Boularias & Chaib-Draa, 2010; Zhang et al., 2019):

$$d_\pi(s, a) = (1 - \gamma)P_0(s)\pi(a|s) + \gamma\pi(a|s) \sum_{\bar{s},\bar{a}} P(s|\bar{s}, \bar{a})d_\pi(s, a).$$

We consider RL in regularized MDPs (Geist et al., 2019), where the policy is optimized with a causal convex regularizer. Mathematically for an MDP $\mathcal{M}^r$ and a strongly convex function $\Omega : \Delta^{\mathcal{A}} \to \mathbb{R}$, the objective in regularized MDPs is to seek π that maximizes the expected discounted sum of rewards, or *return* in short, with policy regularizer Ω:

$$\arg\max_{\pi \in \Delta^{\mathcal{A}}_{\mathcal{S}}} J_\Omega(r, \pi) := \mathbb{E}_\pi \left[\sum_{i=0}^{\infty} \gamma^i \{r(s_i, a_i) - \Omega(\pi(\cdot|s_i))\} \right] = \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim d_\pi} \left[r(s, a) - \Omega(\pi(\cdot|s)) \right]. \tag{1}$$

It turns out that the optimal solution of Eq.(1) is unique (Geist et al., 2019), whereas multiple optimal policies may exist in unregularized MDPs. In later work (Yang et al., 2019), $\Omega(\pi) = -\lambda \mathbb{E}_{a\sim\pi}\phi(\pi(a)), \pi \in \mathcal{D}^{\mathcal{A}}$ was considered for $\lambda > 0$ and $\phi : (0, 1] \to \mathbb{R}$ satisfying some mild conditions. For example, RL with Shannon entropy regularization (Haarnoja et al., 2018) can be recovered by $\phi(\pi(a|s)) = -\log \pi(a|s)$, while RL with Tsallis entropy regularization (Lee et al., 2020) can be recovered from $\phi(\pi(a|s)) = \frac{k}{q-1}(1 - \pi(a|s)^{q-1})$ for $k > 0, q > 1$. The optimal policy π^* for Eq.(1) with Ω from Yang et al. (2019) is shown to be

$$\pi^*(a|s) = \max \left\{ g_\phi \left(\frac{\mu^*(s) - Q^*(s, a)}{\lambda} \right), 0 \right\}, \tag{2}$$

$$Q^*(s, a) = r(s, a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} V^*(s'), V^*(s) = \mu^*(s) - \lambda \sum_{a\in\mathcal{A}} \pi^*(a|s)^2 \phi'(\pi^*(a|s)), \tag{3}$$

where $\phi'(x) = \frac{\partial}{\partial x}\phi(x)$, g_ϕ is an inverse function of f'_ϕ for $f_\phi(x) := x\phi(x), x \in (0, 1]$, and μ^* is a normalization term such that $\sum_{a\in\mathcal{A}} \pi^*(a|s) = 1$. Depending on λ and ϕ, the sparse optimal policy can be derived in Eq.(3) where sparsity is defined as having zero probability for some actions. In contrast, dense optimal policies can only be obtained with Shannon entropy regularization and bounded rewards.

Inverse Reinforcement Learning Given a set of demonstrations from an expert's policy π_E, IRL (Russell, 1998; Ng et al., 2000) is the problem of seeking a reward function from which we can

recover π_E through RL. IRL in unregularized MDPs has however been shown to be an ill-defined problem since *(1)* any constant reward function can rationalize every expert and *(2)* multiple rewards meet the criteria of being a solution (Ng et al., 2000). Maximum entropy IRL (MaxEntIRL) (Ziebart et al., 2008; Ziebart, 2010) is capable of solving the first issue by seeking a reward function that maximizes expert's return along with Shannon entropy of expert's policy. Mathematically for the RL objective J_Ω in Eq.(1) and $\Omega = -\mathcal{H}$ for negative Shannon entropy $\mathcal{H}(\pi) = \mathbb{E}_{a\sim\pi}[-\log \pi(a)]$ (Ho & Ermon, 2016), the objective of MaxEntIRL is

$$\text{MaxEntIRL}(\pi_E) := \underset{r\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}}{\arg\max} \left\{ J_{-\mathcal{H}}(r, \pi_E) - \max_{\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}} J_{-\mathcal{H}}(r, \pi) \right\}. \tag{4}$$

Another commonly used IRL method is Adversarial Inverse Reinforcement Learning (AIRL) (Fu et al., 2018) which involves generative adversarial training (Goodfellow et al., 2014; Ho & Ermon, 2016) to acquire a solution of MaxEntIRL. AIRL considers the structured discriminator (Finn et al., 2016a) $D(s, a) = \sigma(r(s, a) - \log \pi(a|s)) = \frac{e^{r(s,a)}}{e^{r(s,a)}+\pi(a|s)}$ for $\sigma(x) := 1/(1 + e^{-x})$ and iteratively optimizes the following objective:

$$\max_{r\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}} \mathbb{E}_{(s,a)\sim d_{\pi_E}} \left[\log D_{r,\pi}(s, a)\right] + \mathbb{E}_{(s,a)\sim d_\pi} \left[\log(1 - D_{r,\pi}(s, a))\right],$$
$$\max_{\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}} \mathbb{E}_{(s,a)\sim d_\pi} \left[\log D_{r,\pi}(s, a) - \log(1 - D_{r,\pi}(s, a))\right] = \max_{\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}} \mathbb{E}_{(s,a)\sim d_\pi} \left[r(s, a) - \log \pi(a|s)\right]. \tag{5}$$

It turns out that AIRL minimizes the divergence between visitation distributions d_π and d_{π_E} by solving $\min_{\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}} \text{KL}(d_\pi||d_{\pi_E})$ for Kullback-Leibler (KL) divergence KL (Ghasemipour et al., 2019) .

3 Inverse Reinforcement Learning in Regularized MDPs

In this section, we propose the solution of IRL in regularized MDPs and relevant properties in Section 3.1. We then discuss a specific instance of our proposed solution where Tsallis entropy regularizers and multi-variate Gaussian policies are used in continuous action spaces.

3.1 Solutions of regularized IRL

We consider regularized IRL that generalizes MaxEntIRL in Eq.(4) to IRL with a general class of convex policy regularizers:

$$\text{IRL}_\Omega(\pi_E) := \underset{r\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}}{\arg\max} \left\{ J_\Omega(r, \pi_E) - \max_{\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}} J_\Omega(r, \pi) \right\}. \tag{6}$$

For any convex policy regularizer Ω, regularized IRL does not suffer from degenerate solutions since there is a unique optimal policy in any regularized MDP (Geist et al., 2019). While Geist et al. (2019) proposed functional-form solutions of regularized IRL – which is intractable in general – we propose tractable solutions in the following lemma. Our solution is motivated from figuring out a reward function that is capable of converting regularized RL into equivalent divergence minimization problem between π and π_E:

Lemma 1. *For a policy regularizer $\Omega : \Delta^{\mathcal{A}} \to \mathbb{R}$, let us define*

$$t(s, a; \pi) := \Omega'(s, a; \pi) - \mathbb{E}_{a'\sim\pi(\cdot|s)}[\Omega'(s, a'; \pi)] + \Omega(\pi(\cdot|s)) \tag{7}$$

for $\Omega'(s, \cdot; \pi) := \nabla\Omega(\pi(\cdot|s)) := [\nabla_x\Omega(x)]_{x=\pi(\cdot|s)} \in \mathbb{R}^{\mathcal{A}}, s \in \mathcal{S}$. Then, $t(s, a; \pi_E)$ for expert's policy π_E is a solution of regularized IRL with Ω.

Proof. (Abbreviated. See Appendix A for full version.) With $r(s, a) = t(s, a; \pi_E)$, the RL objective in Eq.(1) becomes equivalent to *a problem of minimizing the discounted sum of Bregman divergences between π and π_E*

$$\underset{\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}}{\arg\min}\, \mathbb{E}_\pi \left[\sum_{i=0}^{\infty} \gamma^i D_\Omega^{\mathcal{A}}(\pi(\cdot|s_i)||\pi_E(\cdot|s_i)) \right], \tag{8}$$

where $D_\Omega^{\mathcal{A}}$ is the Bregman divergence (Bregman, 1967) defined by $D_\Omega^{\mathcal{A}}(\pi_1||\pi_2) = \Omega(\pi_1) - \Omega(\pi_2) - \langle \nabla\Omega(\pi_2), \pi_1 - \pi_2 \rangle_{\mathcal{A}}$ for $\pi_1, \pi_2 \in \Delta^{\mathcal{A}}$. Due to the non-negativity of the Bregman divergence, $\pi = \pi_{_E}$ is a solution of Eq.(8) and is unique since Eq.(1) has the unique solution for arbitrary reward functions (Geist et al., 2019). □

Especially for any policy regularizer Ω represented with an expectation over the policy (Yang et al., 2019), **Lemma 1** can be reduced to the following solution in **Corollary 1**:

Corollary 1. *(See Appendix A for proof). For $\Omega(\pi) = -\lambda \mathbb{E}_{a\sim\pi}\phi(\pi(a))$ with $\pi \in \Delta^{\mathcal{A}}$ (Yang et al., 2019) , Eq.(7) becomes*

$$t(s, a; \pi) = -\lambda \cdot \left\{ f'_\phi(\pi(a|s)) - \mathbb{E}_{a'\sim\pi(\cdot|s)}[f'_\phi(\pi(a'|s)) - \phi(\pi(a'|s))] \right\} \tag{9}$$

for $f'_\phi(x) = \frac{\partial}{\partial x}(x\phi(x))$.

Throughout the paper, we denote **reward baseline** by the expectation $\mathbb{E}_{a\sim\pi(\cdot|s)}[f'_\phi(\pi(a|s)) - \phi(\pi(a|s))]$. For *continuous control tasks* with $\Omega(\pi) = -\lambda\mathbb{E}_{a\sim\pi}\phi(\pi(a))$, we can obtain the same form of the reward in Eq.(9) (The proof is in Appendix C). When $\lambda = 1$ and $\phi(x) = -\log x$, it can be shown that $t(s, a; \pi) = \log \pi(a|s)$, which was used as a reward objective in the previous work (Fu et al., 2018), and that the Bregman divergence in Eq.(8) becomes the KL divergence $\mathrm{KL}(\pi(\cdot|s)||\pi_{_E}(\cdot|s))$. Additionally, other solutions of IRL can be found by shaping $t(s, a; \pi_{_E})$ as stated in the following lemma:

Lemma 2. *Let π^* be the solution of Eq.(1) in a regularized MDP $\mathcal{M}^r$ with a regularizer $\Omega : \Delta^{\mathcal{A}} \to \mathbb{R}$ and an arbitrary reward function $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. Then, rewards $r(s, a) - \Phi(s)$ with $\Phi : \mathcal{S} \to \mathbb{R}$ does not change the solution of Eq.(1). The proof is in Section D.4. of Geist et al. (2019).*

From **Lemma 1** and **Lemma 2**, we prove the sufficient condition of rewards being solutions of the IRL problem. However, the necessary condition – a set of those solutions are the only possible solutions – is not proved. We also check how the solution of the IRL problem is related to the normalized state-visitation distribution which can be discussed in the line of distribution matching perspective on imitation learning problems (Ho & Ermon, 2016; Fu et al., 2018; Ghasemipour et al., 2019):

Lemma 3. *Given the policy regularizer Ω, let us define $\bar{\Omega}(d) := \mathbb{E}_{(s,a)\sim d}[\Omega(\bar{\pi}_d(\cdot|s))]$ for an arbitrary normalized state-action visitation distribution $d \in \Delta^{\mathcal{S}\times\mathcal{A}}$ and the policy $\bar{\pi}_d(a|s) := \frac{d(s,a)}{\sum_{a'\in\mathcal{A}} d(s,a')}$ induced by d. Then, Eq.(7) is equal to*

$$t(s, a; \bar{\pi}_d) = [\nabla\bar{\Omega}(d)](s, a). \tag{10}$$

When $\bar{\Omega}(d)$ is strictly convex and a solution $t(s, a; \pi_{_E}) = \nabla\bar{\Omega}(d_{\pi_{_E}})$ of IRL in Eq.(10) is used, the RL objective in $Eq.(1)$ is equal to

$$\arg\min_{\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}} D_{\bar{\Omega}}^{\mathcal{S}\times\mathcal{A}}(d_\pi||d_{\pi_{_E}}),$$

where $D_{\bar{\Omega}}^{\mathcal{S}\times\mathcal{A}}$ is a Bregman divergence among visitation distributions defined by $D_{\bar{\Omega}}^{\mathcal{S}\times\mathcal{A}}(d_1||d_2) = \bar{\Omega}(d_1) - \bar{\Omega}(d_2) - \langle\nabla\bar{\Omega}(d_2), d_1 - d_2\rangle$ for visitation distributions d_1 and d_2. The proof is in Appendix D.

Here, although the strict convexity of a policy regularizer Ω does not guarantee our assumption on the strict convexity of $\bar{\Omega}$, it has been shown to be true for Shannon (Ho & Ermon, 2016) and Tsallis entropy regularizations (Lee et al., 2018).

3.2 IRL with Tsallis entropy regularization and Gaussian policies

For continuous controls, multi-variate Gaussian policies are widely used in practice (Schulman et al., 2015; 2017) and we consider IRL problems with those policies in this subsection. In particular, we consider IRL with Tsallis entropy regularizer $\Omega(\pi(\cdot|s)) = -\mathcal{T}_q^k(\pi(\cdot|s)) = -\mathbb{E}_{a\sim\pi(\cdot|s)}[\frac{k}{q-1}(1 - \pi(a|s)^{q-1})]$ (Lee et al., 2018; Yang et al., 2019; Lee et al., 2020) for a multi-variate Gaussian policy $\pi(\cdot|s) = \mathcal{N}(\boldsymbol{\mu}(s), \boldsymbol{\Sigma}(s))$ with $\boldsymbol{\mu}(s) = [\mu_1(s), ..., \mu_d(s)]^T$, $\boldsymbol{\Sigma}(s) = \mathrm{diag}\{(\sigma_1(s))^2, ..., (\sigma_d(s))^2\}$. In such a case, we can obtain analytical forms of the following quantities:

Tsallis entropy. The analytical form of Tsallis entropy for a multi-variate Gaussian policy is

$$\mathcal{T}_q^k(\pi(\cdot|s)) = \frac{k(1 - e^{(1-q)\mathcal{R}_q(\pi(\cdot|s))})}{q-1}, \mathcal{R}_q(\pi(\cdot|s)) = \sum_{i=1}^{d} \left\{ \log(\sqrt{2\pi}\sigma_i(s)) - \frac{\log q}{2(1-q)} \right\}$$

for Renyi entropy $\mathcal{R}_q$. Its derivation is given in Appendix F.

Reward baseline. In general, the reward baseline term $\mathbb{E}_{a\sim\pi(\cdot|s)}[f'_\phi(\pi(a|s)) - \phi(\pi(a|s))]$ in **Corollary 1** is generally intractable except for either discrete control problems or when Shannon entropy regularization is considered – where the reward baseline is equal to -1. Interestingly, as long as the analytical form of Tsallis entropy can be derived, that of the corresponding reward baseline can also be derived since the reward baseline satisfies

$$\mathbb{E}_{a\sim\pi(\cdot|s)}[f'_\phi(\pi(a|s)) - \phi(\pi(a|s))] = (q-1)\mathbb{E}_{a\sim\pi(\cdot|s)}[\phi(\pi(a|s)] - k = (q-1)\mathcal{T}_q^k(\pi(\cdot|s)) - k.$$

Here, the first equality holds with $f'_\phi(x) = \frac{k}{q-1}(1 - qx^{q-1}) = q\phi(x) - k$ for Tsallis entropy regularization.

Bregman divergence associated with Tsallis entropy. For two different multivariate Gaussian policies, we derive the analytical form of the Bregman divergence (associated with Tsallis entropy) between two policies. The resultant divergence has a complicated form, so we leave it in Appendix F.3 with its derivation.

For deeper understanding of Tsallis entropy and its associated Bregman divergence in regularized MDPs, we consider an example in Figure 1. We first assume that both learning agents' and experts' policies follow uni-variate Gaussian distributions $\pi = \mathcal{N}(\mu, \sigma^2)$ and $\pi_E = \mathcal{N}(0, (e^{-3})^2)$, respectively. We then evaluate the Bregman divergence by using the analytical form of the above quantities in Figure 1 by varying q from 1.0 – which corresponds to the KL divergence – to 2.0. We observe that the constant q from the Tsallis entropy affects the sensitivity of the associated Bregman divergence w.r.t. the mean and standard deviation of the learning agent's policy π. Specifically, as q increases, the size of the valley – relatively red regime in Figure 1 – across the μ-axis and $\log\sigma$-axis decreases. This suggests that for larger q, minimizing the Bregman divergence requires more tightly matching means and variances of π and π_E.

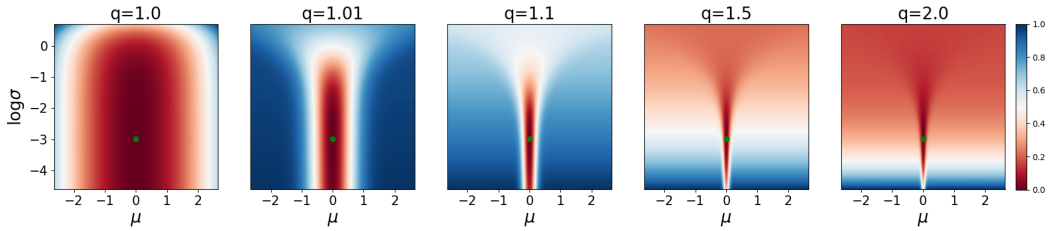


Figure 1: Bregman divergence $D_\Omega(\pi||\pi_E)$ associated with Tsallis entropy ($\Omega(\pi) = -\mathcal{T}_q^1(\pi)$) between two uni-variate Gaussian distributions $\pi = \mathcal{N}(\mu, \sigma^2)$ and $\pi_E = \mathcal{N}(0, (e^{-3})^2)$ (green point in each subplot). In each subplot, we normalized the Bregman divergence so that the maximum value becomes 1. Note that for $q = 1$, $D_\Omega(\pi||\pi_E)$ becomes the KL divergence $\mathrm{KL}(\pi||\pi_E)$.

4 Algorithmic Consideration

Although we derive a solution for IRL problems with a general policy regularizer Ω in **Lemma 1**, we focus on developing an IRL algorithm for $\Omega(\pi) = -\lambda\mathbb{E}_{a\sim\pi}[\phi(\pi(a))]$ (Yang et al., 2019). To recover the reward function $t(s, a; \pi_E)$ in **Corollary 1**, we design an adversarial training objective as follows. Similar to AIRL (Fu et al., 2018), we consider the following structured discriminator associated with π, r and t in **Lemma 1**:

$$D_{r,\pi}(s, a) = \sigma(r(s, a) - t(s, a; \pi)), \sigma(z) = \frac{1}{1 + e^{-z}}, z \in \mathbb{R}.$$

Note that we can recover the discriminator of AIRL in Eq.(5) when $t(s, a) = \log\pi(a|s)$ ($\phi(x) = \log x$ and $\lambda = 1$). Then, we consider the following optimization objective of the discriminator which

is the same as that of AIRL:

$$\hat{t}(s,a;\pi) := \underset{r\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}}{\arg\max}\ \mathbb{E}_{(s,a)\sim d_{\pi_E}}\left[\log D_{r,\pi}(s,a)\right] + \mathbb{E}_{(s,a)\sim d_{\pi}}\left[\log(1-D_{r,\pi}(s,a))\right]. \tag{11}$$

Since the function $x \mapsto a\log\sigma(x) + b\log(1-\sigma(x))$ attains its maximum at $\sigma(x) = \frac{a}{a+b}$, or equivalently at $x = \log\frac{a}{b}$ (Goodfellow et al., 2014; Mescheder et al., 2017), it can be shown that

$$\hat{t}(s,a;\pi) = t(s,a;\pi) + \log\frac{d_{\pi_E}(s,a)}{d_{\pi}(s,a)}. \tag{12}$$

When $\pi = \pi_E$ in Eq.(12), we have $\hat{t}(s,a;\pi_E) = t(s,a;\pi_E)$ since $d_\pi = d_{\pi_E}$, which means the maximizer $\hat{t}$ becomes the solution of IRL after the agent successfully imitates the expert's policy π_E. To do so, we consider the following iterative algorithm. Assuming that we find out the optimal $\hat{t}(s,a;\pi^{(i)})$ in Eq.(12) for the policy $\pi^{(i)}$ of the i-th iteration, we get the policy $\pi^{(i+1)}$ by optimizing the following objective:

$$\pi^{(i+1)} := \underset{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}}{\arg\max}\ \mathbb{E}_{(s,a)\sim d_{\pi}}\left[\hat{t}(s,a;\pi^{(i)}) - \Omega(\pi(\cdot|s))\right]. \tag{13}$$

Note that the expectation in Eq.(13) can be decomposed into the following two terms:

$$\begin{aligned}\mathbb{E}_{(s,a)\sim d_{\pi}}\left[\hat{t}(s,a;\pi^{(i)}) - \Omega(\pi(\cdot|s))\right] &= \mathbb{E}_{(s,a)\sim d_{\pi}}\left[t(s,a;\pi^{(i)}) - \Omega(\pi(\cdot|s))\right] - \mathrm{KL}(d_\pi||d_{\pi_E}) \\ &= -\underbrace{\mathbb{E}_{(s,a)\sim d_{\pi}}\left[D_{\Omega}^{\mathcal{A}}(\pi(\cdot|s)||\pi^{(i)}(\cdot|s))\right]}_{\text{(I)}} - \underbrace{\mathrm{KL}(d_\pi||d_{\pi_E})}_{\text{(II)}},\end{aligned} \tag{14}$$

where the second equality follows since **Lemma 1** shows us that $t(s,a;\pi^{(i)})$ is a reward function that makes $\pi^{(i)}$ an optimal policy in the Ω-regularized MDP. Minimizing term (II) in Eq.(14) makes $\pi^{(i+1)}$ close to π_E while minimizing term (I) can be regarded as a conservative policy optimization around the policy $\pi^{(i)}$ (Schulman et al., 2015).

In practice, we parameterize our reward and policy approximations with neural networks and train them using an off-policy Regularized Actor-Critic (RAC) (Yang et al., 2019) as described in **Algorithm 1** in Appendix G. We evaluate our Regularized Adversarial Inverse Reinforcement Learning (RAIRL) approach across various scenarios, below.

5 Experiments

We summarize the experimental setup as follows. In our experiments, we consider $\Omega(\pi) = -\lambda\mathbb{E}_{a\sim\pi}[\phi(\pi(a)]$ with the following regularizers from Yang et al. (2019): *(1) Shannon entropy* ($\phi(x) = -\log x$), *(2) Tsallis entropy regularizer* ($\phi(x) = \frac{k}{q-1}(1-x^{q-1})$), *(3)* exp *regularizer* ($\phi(x) = e - e^x$), *(4)* cos *regularizer* ($\phi(x) = \cos(\frac{\pi}{2}x)$), *(5)* sin *regularizer* ($\phi(x) = 1 - \sin\frac{\pi}{2}x$). In addition, we model the reward approximator in **Algorithm 1** as a neural network with either one of the following models: *(1) Non-structured model (NSM)* – a simple feed-forward neural network that outputs real values used in AIRL (Fu et al., 2018) – and *(2) Density-based model (DBM)* (Figure H.2 in Appendix) – a model using a neural network for π (softmax for discrete controls and multi-variate Gaussian model for continuous controls) of the solution in Eq.(1). Other settings are summarized in Appendix H. For all experiments, we use 5 runs and report 95% confidence interval.

5.1 Experiment 1: Multi-armed Bandit (Discrete Action)

We consider a 4-armed bandit environment as shown in Figure 2 (left). An expert's policy π_E is assumed to be either *dense* (with probability 0.1, 0.2, 0.3, 0.4 for $a = 0,1,2,3$) or *sparse* (with probability 0, 0, 1/3, 2/3 for $a = 0,1,2,3$). For those experts, we use RAIRL with actions sampled from π_E and compare learned rewards with the ground truth reward $t(s,a;\pi_E)$ in **Lemma 1**. When π_E is dense, RAIRL successfully acquires the ground truth rewards irrespective of the reward model choices. When sparse π_E is used, however, RAIRL with a non-structured model (RAIRL-NSM)

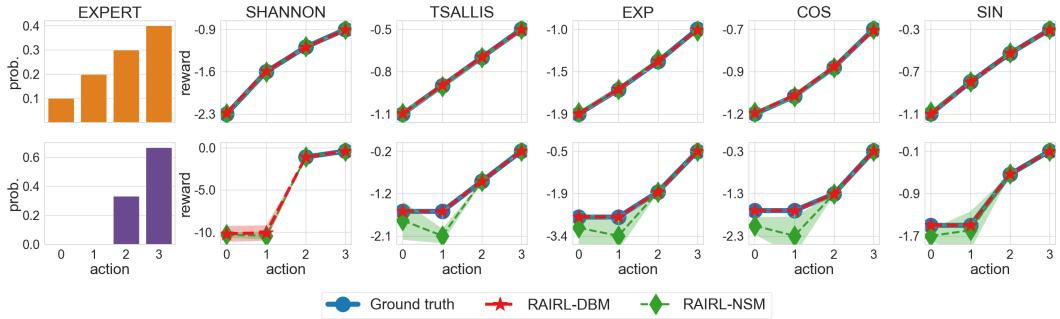


Figure 2: Expert Policy (*Left*). Reward learned by RAIRL with different types of policy regularizers (*Right*). Either one of dense (*Top row*) or sparse (*Bottom row*) expert policies π_E is considered.

failed to recover the rewards for $a = 0, 1$ – where $\pi_E(a) = 0$ – due to the lack of samples at the end of the imitation. On the other hand, RAIRL with a density-based model (RAIRL-DBM) can recover the correct rewards due to the softmax layer which maintains the sum over the outputs equal to 1. Therefore, we argue that using DBM is necessary for correct reward acquisition since a set of demonstration is generally sparse. In the following experiment, we show the choice of reward models indeed affects the performance of rewards.

5.2 Experiment 2: Bermuda World (Continuous Observation, Discrete Action)

We consider an environment with a 2-dimensional continuous state space as described in Figure 3. At each episode, the learning agent is initialized uniformly on the x-axis between -5 and 5, and there are 8 possible actions – an angle in $\{-\pi, -\frac{3}{4}\pi, ..., \frac{1}{2}\pi, \frac{3}{4}\pi\}$ that determines the direction of movement. An expert in Bermuda World considers 3 target positions $(-5, 10), (0, 10), (5, 10)$ and behaves stochastically as shown in Figure 3 (*Left*). With 1000 demonstrations sampled from the expert, the agent is evaluated during training with the Bregman divergences that correspond to each policy regularizer. RAIRL-DBM is shown to minimize the target divergence more effectively compared to RAIRL-NSM although both achieve comparable performances during training. Moreover, we substitute λ with $1, 5, 10$ and observe that learning with λ larger than 1 returns better rewards – only $\lambda = 1$ was considered in AIRL (Fu et al., 2018). Note that in all cases, the minimum divergence achieved by RAIRL is comparable with that of behavioral cloning (BC). This is because BC performs sufficiently well when sufficiently many demonstration is given. We think the divergence of BC may be the near-optimal divergence that can be achieved with our policy neural network model.

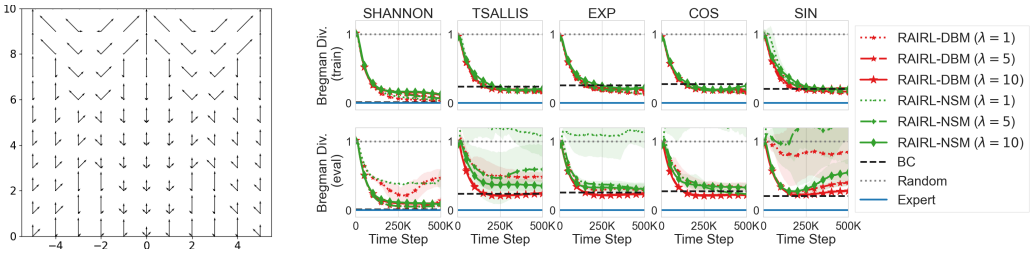


Figure 3: An expert's policy π_E (*Left*), averaged Bregman divergence during training (*Right, Top row*) and the divergence during evaluation (*Right, Bottom row*) in Bermuda World. In each column of the figure on the right, different policy regularizers and their respective target divergences are considered. The results are reported after normalization with the divergence of uniform random policy, and that of behavioral cloning (BC) is reported for comparison.

5.3 Experiment 3: MuJoCo (Continuous Observation and Action)

We evaluate RAIRL on MuJoCo continuous control tasks (*Hopper-v2*, *Walker-v2*, *HalfCheetah-v2*, *Ant-v2*) as follows. We assume multivariate-Gaussian policies (with diagonal covairance matrices)

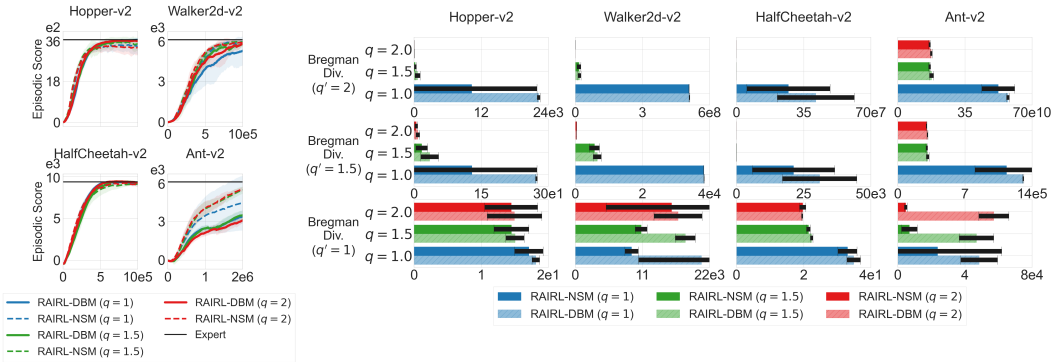


Figure 4: Averaged episodic score (*Left*) and Bregman divergences with Tsallis entropy $\mathcal{T}_{q'}^1$ with $q' = 1, 1.5, 2$ (*Right*) during training in MuJoCo environments. RAIRL with $\mathcal{T}_q^1$ regularizer with $q = 1, 1.5, 2$ is considered.

for both learner's policy π and expert's policy π_E. Instead of $\tanh$-squashed policy in Soft-Actor Critic (Haarnoja et al., 2018), we use hyperbolized environments – where $\tanh$ is regarded as a part of the environment – with additional engineering on the policy networks (See Appendix H.4 for details). We use 100 demonstrations from π_E to evaluate RAIRL. In MuJoCo experiments, we focus on a set of Tsallis entropy regularizer ($\Omega(\pi) = -\mathcal{T}_q^1(\pi)$) with $q = 1, 1.5, 2$ – where Tsallis entropy becomes Shannon entropy for $q = 1$. We then exploit the analytical quantities for multi-variate Gaussian distributions in Section 3.2 to stabilize RAIRL and evaluate its performance in terms of various Bregman divergences.

The evaluation performances for RAIRL are depicted in Figure 4. We evaluate π with an episodic score (Figure 4, *Left*) and three Bregman divergences (Figure 4, *Right*) $\Omega(\pi) = -\mathcal{T}_{q'}^1(\pi)$ with $q' = 1, 1.5, 2$, respectively. Note that the objective of RAIRL with $\Omega = -\mathcal{T}_q$ is to minimize the corresponding Bregman divergence with $q' = q$. In Figure 4 (*Left*), both RAIRL-DBM and RAIRL-NSM are shown to achieve the expert performance, irrespective of q, in *Hopper-v2*, *Walker-v2*, and *HalfCheetah-v2*. In contrast, RAIRL in *Ant-v2* fails to achieve the expert's performance within 2,000,000 steps and RAIRL-NSM highly outperforms RAIRL-DBM in our setting. Although the evaluation scores are comparable for all methods in *Hopper-v2*, *Walker-v2*, and *HalfCheetah-v2*, respective divergences are shown to be highly different from one another as shown in Figure 4 (*Right*). RAIRL with $q = 2$ in most cases achieves the minimum Bregman divergence (for all three divergences with $q' = 1, 1.5, 2$), whereas RAIRL with $q = 1$ – which corresponds to AIRL (Fu et al., 2018) – achieves the maximum divergence in most cases. This result is in alignment with our intuition from Section 3.2; as q increases, minimizing the Bregman divergence requires much tighter matching between π and π_E. Therefore, we believe RAIRL is suitable for making imitation learning safer in practice. Unfortunately, RAIRL fails to acquire a reward function that effectively minimizes the target divergence in continuous controls. We believe this is because π is a probability density function in continuous controls and causes large variance during training, while π is a mass function and is well-bounded in discrete control problems.

6 Discussion and Future Works

We consider the problem of IRL in regularized MDPs (Geist et al., 2019), assuming a general class of convex policy regularizers. We theoretically derive its solution (a set of reward functions) and show that learning with these rewards is equivalent to a specific instance of imitation learning – i.e., one that minimizes the Bregman divergence associated with policy regularizers. We propose RAIRL – a practical sampled-based IRL algorithm in regularized MDPs – and evaluate its applicability on policy imitation (for discrete and continuous controls) and reward acquisition (for discrete control).

Finally, recent advances in imitation learning and IRL are built from the perspective of regarding imitation learning as statistical divergence minimization problems (Ghasemipour et al., 2019). Although Bregman divergence is known to cover various divergences, it does not include some divergence families such as f-divergence (Csiszár, 1963; Amari, 2009). Therefore, we believe that considering RL with policy regularization different from Geist et al. (2019) and its inverse problem is a possible way of finding the links between imitation learning and various statistical distances.

REFERENCES

Shun-Ichi Amari. α-divergence is unique, belonging to both f-divergence and bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, 2009.

Abdeslam Boularias and Brahim Chaib-Draa. Bootstrapping apprenticeship learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 289–297, 2010.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutató Int. Közl*, 8:85–108, 1963.

Robert Dadashi, Léonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal wasserstein imitation learning. *arXiv preprint arXiv:2006.04678*, 2020.

Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*, 2016a.

Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 49–58, 2016b.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1582–1591, 2018.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 2160–2169, 2019.

Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In *Proceedings of the 3rd Conference on Robot Learning (CoRL)*, 2019.

Ian Goodfellow, Jean Pouget Abadie, Mehdi Mirza, Bing Xu, David Warde Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2672–2680, 2014.

Xin Guo, Johnny Hong, and Nan Yang. Ambiguity set and learning via Bregman and Wasserstein. *arXiv preprint arXiv:1705.08056*, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1861–1870, 2018.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4565–4573, 2016.

Lee K Jones and Charles L Byrne. General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE transactions on Information Theory*, 36(1):23–30, 1990.

Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Maximum causal tsallis entropy imitation learning. In *Advances in Neural Information Processing Systems*, pp. 4403–4413, 2018.

Kyungjae Lee, Sungyub Kim, Sungbin Lim, Sungjoon Choi, Mineui Hong, Jaein Kim, Yong-Lae Park, and Songhwai Oh. Generalized Tsallis entropy reinforcement learning and its application to soft mobile robots. Robotics: Science and Systems Foundation, 2020.

Kun Li, Mrinal Rath, and Joel W Burdick. Inverse reinforcement learning via function approximation for clinical motion analysis. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 610–617. IEEE, 2018.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 2391–2400, 2017.

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

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1928–1937, 2016.

Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pp. 663–670, 2000.

Frank Nielsen and Richard Nock. On Rényi and Tsallis entropies and divergences for exponential families. *arXiv preprint arXiv:1105.3259*, 2011.

Stuart Russell. Learning agents for uncertain environments. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp. 101–103, 1998.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 1889–1897, 2015.

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

Sahand Sharifzadeh, Ioannis Chiotellis, Rudolph Triebel, and Daniel Cremers. Learning to drive using inverse reinforcement learning and deep Q-networks. *arXiv preprint arXiv:1612.03653*, 2016.

Umar Syed, Michael Bowling, and Robert E Schapire. Apprenticeship learning using linear programming. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pp. 1032–1039, 2008.

Zheng Wu, Liting Sun, Wei Zhan, Chenyu Yang, and Masayoshi Tomizuka. Efficient sampling-based maximum entropy inverse reinforcement learning with application to autonomous driving. *IEEE Robotics and Automation Letters*, 5(4):5355–5362, 2020.

Wenhao Yang, Xiang Li, and Zhihua Zhang. A regularized approach to sparse optimal policy in reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5938–5948, 2019.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. GenDice: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2019.

Brian D Ziebart. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. 2010.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, volume 8, pp. 1433–1438, 2008.

A PROOF OF LEMMA 1

Let us define $\pi^s = \pi(\cdot|s)$. For $r(s,a) = t(s,a;\pi_{_E})$, the RL objective Eq.(1) satisfies

$$
\begin{aligned}
&\mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\left\{r(s_i,a_i)-\Omega(\pi^{s_i})\right\}\right] \overset{\text{(i)}}{=} \mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\left\{t(s,a;\pi_{_E})-\Omega(\pi^{s_i})\right\}\right] \\
&\overset{\text{(ii)}}{=} \mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\left\{\Omega'(s_i,a_i;\pi_{_E})-\mathbb{E}_{a\sim\pi_{_E}^{s_i}}\Omega'(s_i,a;\pi_{_E})+\Omega(\pi_{_E}^{s_i})-\Omega(\pi^{s_i})\right\}\right] \\
&\overset{\text{(iii)}}{=} \mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\left\{\mathbb{E}_{a\sim\pi^{s_i}}\Omega'(s_i,a;\pi_{_E})-\mathbb{E}_{a\sim\pi_{_E}^{s_i}}\Omega'(s_i,a;\pi_{_E})+\Omega(\pi_{_E}^{s_i})-\Omega(\pi^{s_i})\right\}\right] \\
&\overset{\text{(iv)}}{=} \mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\left\{\langle\nabla\Omega(\pi_{_E}^{s_i}),\pi^{s_i}\rangle_{\mathcal{A}}-\langle\nabla\Omega(\pi_{_E}^{s_i}),\pi_{_E}^{s_i}\rangle_{\mathcal{A}}+\Omega(\pi_{_E}^{s_i})-\Omega(\pi^{s_i})\right\}\right] \\
&= \mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\left\{\Omega(\pi_{_E}^{s_i})-\Omega(\pi^{s_i})+\langle\nabla\Omega(\pi_{_E}^{s_i}),\pi^{s_i}\rangle_{\mathcal{A}}-\langle\nabla\Omega(\pi_{_E}^{s_i}),\pi_{_E}^{s_i}\rangle_{\mathcal{A}}\right\}\right] \\
&= -\mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\left\{\Omega(\pi^{s_i})-\Omega(\pi_{_E}^{s_i})-\langle\nabla\Omega(\pi_{_E}^{s_i}),\pi^{s_i}\rangle_{\mathcal{A}}+\langle\nabla\Omega(\pi_{_E}^{s_i}),\pi_{_E}^{s_i}\rangle_{\mathcal{A}}\right\}\right] \\
&= -\mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\left\{\Omega(\pi^{s_i})-\Omega(\pi_{_E}^{s_i})-\langle\nabla\Omega(\pi_{_E}^{s_i}),\pi^{s_i}-\pi_{_E}^{s_i}\rangle_{\mathcal{A}}\right\}\right] \overset{\text{(v)}}{=} -\mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i D_\Omega^{\mathcal{A}}(\pi^{s_i}||\pi_{_E}^{s_i}))\right],
\end{aligned}
\tag{15}
$$

where (i) follows from the assumption $r(s,a) = t(s,a;\pi_{_E})$ in **Lemma 1**, (ii) follows from the definition of $t(s,a;\pi)$ in Eq.(7), (iii) follows since taking the inner expectation first does not change the overall expectation, (iv) follows from the definition of Ω' in **Lemma 1** and $\sum_{a\in\mathcal{A}}\pi(a)[\nabla\Omega(\pi)](a) = \langle\nabla\Omega(\pi),\pi\rangle_{\mathcal{A}}$, and (v) follows from the definition of Bregman divergence, i.e., $D_\Omega^{\mathcal{A}}(\pi_1||\pi_2) = \Omega(\pi_1)-\Omega(\pi_2)-\langle\nabla\Omega(\pi_2),\pi_1-\pi_2\rangle_{\mathcal{A}}$. Due to the non-negativity of $D_\Omega^{\mathcal{A}}$, Eq.(15) is less than or equal to zero which can be achieved when $\pi = \pi_{_E}$.

B PROOF OF COROLLARY 1

Let $a \in \{1,...,|\mathcal{A}|\}$ and $\pi_a = \pi(a)$ for simplicity. For

$$
\Omega(\pi) = -\lambda\mathbb{E}_{a\sim\pi}\phi(\pi_a) = -\lambda\sum_{a\in\mathcal{A}}\pi_a\phi(\pi_a) = -\lambda\sum_{a\in\mathcal{A}}f_\phi(\pi_a)
$$

with $f_\phi(x) = x\phi(x)$, we have

$$
\nabla\Omega(\pi) = -\lambda\nabla_{\pi_1,...,\pi_{|\mathcal{A}|}}\sum_{a\in\mathcal{A}}f_\phi(\pi_a) = -\lambda[f'_\phi(\pi_1),...,f'_\phi(\pi_{|\mathcal{A}|})]^T
$$

for $f'_\phi(x) = \frac{\partial}{\partial x}(x\phi(x))$. Therefore, for $\pi^s = \pi(\cdot|s)$ we have

$$
\begin{aligned}
t(s,a;\pi) &= [\nabla\Omega(\pi^s)](a) - \langle\nabla\Omega(\pi^s),\pi^s\rangle_{\mathcal{A}} + \Omega(\pi^s) \\
&= -\lambda f'_\phi(\pi_a^s) - \left(\sum_{a'\in\mathcal{A}}\pi_{a'}^s(-\lambda f'_\phi(\pi_{a'}^s))\right) + \left(-\lambda\sum_{a'\in\mathcal{A}}\pi_{a'}^s\phi(\pi_{a'}^s)\right) \\
&= -\lambda\left\{f'_\phi(\pi_a^s) - \left(\sum_{a'\in\mathcal{A}}\pi_{a'}^s f'_\phi(\pi_{a'}^s)\right) + \left(\sum_{a'\in\mathcal{A}}\pi_{a'}^s\phi(\pi_{a'}^s)\right)\right\} \\
&= -\lambda\left\{f'_\phi(\pi_a^s) - \sum_{a'\in\mathcal{A}}\pi_{a'}^s\left(f'_\phi(\pi_{a'}^s)-\phi(\pi_{a'}^s)\right)\right\} \\
&= -\lambda\left\{f'_\phi(\pi_a^s) - \mathbb{E}_{a'\sim\pi^s}\left[f'_\phi(\pi_{a'}^s)-\phi(\pi_{a'}^s)\right]\right\}.
\end{aligned}
$$

C PROOF OF OPTIMAL REWARDS ON CONTINUOUS CONTROLS

Note that for two continuous distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ having probability density functions $p_1(x)$ and $p_2(x)$, respectively, the Bregman divergence can be defined as (Guo et al., 2017; Jones & Byrne, 1990)

$$D_\omega^{\mathcal{X}}(\mathbb{P}_1||\mathbb{P}_2) := \int_{\mathcal{X}} \{\omega(p_1(x)) - \omega(p_2(x)) - \omega'(p_2(x))(p_1(x) - p_2(x))\}\, dx, \tag{16}$$

where $\omega'(x) := \frac{\partial}{\partial x}\omega(x)$ and the divergence is measure point-wisely on $x \in \mathcal{X}$. Let us assume

$$\Omega(\pi) = \int_{\mathcal{A}} \omega(\pi(a))da \tag{17}$$

for a probability density function π on $\mathcal{A}$ and define

$$t(s, a; \pi) := \omega'(\pi^s(a)) - \int_{\mathcal{A}} [\pi^s(a')\omega'(\pi^s(a')) - \omega(\pi^s(a'))]\, da'. \tag{18}$$

for $\pi^s = \pi(\cdot|s)$. For $r(s, a) = t(s, a; \pi_{_E})$, the RL objective in Eq.(1) satisfies

$$\begin{aligned}
&\mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\left\{r(s_i, a_i) - \Omega(\pi^{s_i})\right\}\right] = \mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\left\{t(s_i, a_i; \pi_{_E}) - \Omega(\pi^{s_i})\right\}\right] \\
&= \mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\left\{\omega'(\pi_{_E}^{s_i}(a_i)) - \int_{\mathcal{A}}\left[\pi_{_E}^{s_i}(a)\omega'(\pi_{_E}^{s_i}(a)) - \omega(\pi_{_E}^{s_i}(a))\right]da - \int_{\mathcal{A}}\omega(\pi^{s_i}(a))da\right\}\right] \\
&= \mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\left\{\int_{\mathcal{A}}\pi^{s_i}(a)\omega'(\pi_{_E}^{s_i}(a))da - \int_{\mathcal{A}}\left[\pi_{_E}^{s_i}(a)\omega'(\pi_{_E}^{s_i}(a)) - \omega(\pi_{_E}^{s_i}(a)) + \omega(\pi^{s_i}(a))\right]da\right\}\right] \\
&= \mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\int_{\mathcal{A}}\left\{\pi^{s_i}(a)\omega'(\pi_{_E}^{s_i}(a)) - \left[\pi_{_E}^{s_i}(a)\omega'(\pi_{_E}^{s_i}(a)) - \omega(\pi_{_E}^{s_i}(a)) + \omega(\pi^{s_i}(a))\right]\right\}da\right] \\
&= \mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\int_{\mathcal{A}}\left\{\omega(\pi_{_E}^{s_i}(a)) - \omega(\pi^{s_i}(a)) + \pi^{s_i}(a)\omega'(\pi_{_E}^{s_i}(a)) - \pi_{_E}^{s_i}(a)\omega'(\pi_{_E}^{s_i}(a))\right\}da\right] \\
&= -\mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\int_{\mathcal{A}}\left\{\omega(\pi^{s_i}(a)) - \omega(\pi_{_E}^{s_i}(a)) - \pi^{s_i}(a)\omega'(\pi_{_E}^{s_i}(a)) + \pi_{_E}^{s_i}(a)\omega'(\pi_{_E}^{s_i}(a))\right\}da\right] \\
&= -\mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i\int_{\mathcal{A}}\left\{\omega(\pi^{s_i}(a)) - \omega(\pi_{_E}^{s_i}(a)) - \omega'(\pi_{_E}^{s_i}(a))\left(\pi^{s_i}(a) - \pi_{_E}^{s_i}(a)\right)\right\}da\right] \\
&= -\mathbb{E}_\pi\left[\sum_{i=0}^{\infty}\gamma^i D_\omega^{\mathcal{A}}(\pi^{s_i}||\pi_{_E}^{s_i}))\right],
\end{aligned} \tag{19}$$

where (i) follows from $r(s, a) = t(s, a; \pi_{_E})$, (ii) follows from Eq.(17) and Eq.(18), and (iii) follows from the definition of Bregman divergence in Eq.(16). Due to the non-negativity of D_ω, Eq.(21) is less than or equal to zero which can be achieved when $\pi = \pi_{_E}$. Also, $\pi = \pi_{_E}$ is a unique solution since Eq.(1) has a unique solution for arbitrary reward functions.

D PROOF OF LEMMA 3

RL objective in Regularized MDPs w.r.t. normalized visitation distributions. For a reward function $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and a strongly convex function $\Omega : \Delta_{\mathcal{S}}^{\mathcal{A}} \to \mathbb{R}$, the RL objetive $J_\Omega(r, \pi)$ in Eq.(1) is equivalent to

$$\arg\max_\pi \bar{J}_{\bar{\Omega}}(r, d_\pi) := \langle r, d_\pi\rangle_{\mathcal{S}\times\mathcal{A}} - \bar{\Omega}(d_\pi), \tag{20}$$

where for a set $\mathcal{D}$ of *normalized* visitation distributions (Syed et al., 2008)

$$\mathcal{D} := \left\{d \in \Delta^{\mathcal{S}\times\mathcal{A}} : \sum_{a'} d(s', a') = (1-\gamma)P_0(s') + \gamma\sum_{s,a}P(s'|s, a)d(s, a), \forall s' \in \mathcal{S}\right\},$$

we define $\bar{\Omega}(d) := \mathbb{E}_{(s,a)\sim d}[\Omega(\bar{\pi}_d(\cdot|s))]$ and $\bar{\pi}_d(\cdot|s) := \frac{d(s,\cdot)}{\sum_{a'} d(s,a')} \in \Delta_{\mathcal{S}}^{\mathcal{A}}$ for $d \in \mathcal{D}$ and use $\bar{\pi}_{d_\pi}(\cdot|s) = \pi(\cdot|s)$ for all $s \in \mathcal{S}$. For $\bar{\Omega} : \mathcal{D} \to \mathbb{R}$, its convex conjugate $\bar{\Omega}^*$ is

$$\begin{aligned}
\bar{\Omega}^*(r) :&= \max_{d\in\mathcal{D}} \bar{J}_{\bar{\Omega}}(r,d) \\
&= \max_{d\in\mathcal{D}} \langle r, d\rangle_{\mathcal{S}\times\mathcal{A}} - \bar{\Omega}(d) \\
&\overset{\text{(i)}}{=} \max_{\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}} \langle r, d_\pi\rangle_{\mathcal{S}\times\mathcal{A}} - \bar{\Omega}(d_\pi) \\
&= \max_{\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}} \sum_{s,a} d_\pi(s,a)\left[r(s,a) - \Omega(\pi(a|s))\right] \\
&= (1-\gamma)\cdot \max_{\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}} J_\Omega(r,\pi),
\end{aligned} \tag{21}$$

where (i) follows from using the one-to-one correspondence between policies and visitation distributions (Syed et al., 2008; Ho & Ermon, 2016). Note that Eq.(21) is equal to the optimal discounted average return in regularized MDPs.

IRL objective in Regularized MDPs w.r.t. normalized visitation distributions. By using the RL objective in Eq.(20), we can rewrite the IRL objective in Eq.(6) w.r.t. the normalized visitation distributions as the maximization of the following objective over $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$:

$$\begin{aligned}
&(1-\gamma)\cdot\left\{J_\Omega(r,\pi_E) - \max_{\pi\in\Delta_{\mathcal{S}}^{\mathcal{A}}} J_\Omega(r,\pi)\right\} \\
&= \bar{J}_{\bar{\Omega}}(r,d_{\pi_E}) - \max_{d\in\mathcal{D}} \bar{J}_{\bar{\Omega}}(r,d) \\
&= \min_{d\in\mathcal{D}} \left\{\bar{J}_{\bar{\Omega}}(r,d_{\pi_E}) - \bar{J}_{\bar{\Omega}}(r,d)\right\} \\
&= \min_{d\in\mathcal{D}} \left\{\left(\langle r, d_{\pi_E}\rangle_{\mathcal{S}\times\mathcal{A}} - \bar{\Omega}(d_{\pi_E})\right) - \left(\langle r, d\rangle_{\mathcal{S}\times\mathcal{A}} - \bar{\Omega}(d)\right)\right\} \\
&= \min_{d\in\mathcal{D}} \left\{\bar{\Omega}(d) - \bar{\Omega}(d_{\pi_E}) - \langle r, d - d_{\pi_E}\rangle_{\mathcal{S}\times\mathcal{A}}\right\}.
\end{aligned} \tag{22}$$

Note that if $\nabla\bar{\Omega}(d)$ is well-defined and $r = \nabla\bar{\Omega}(d_{\pi_E})$ for any strictly convex $\bar{\Omega}$, Eq.(22) is equal to

$$\min_{d\in\mathcal{D}} \left\{\bar{\Omega}(d) - \bar{\Omega}(d_{\pi_E}) - \langle \nabla\bar{\Omega}(d_{\pi_E}), d - d_{\pi_E}\rangle_{\mathcal{S}\times\mathcal{A}}\right\} = \min_{d\in\mathcal{D}} D_{\bar{\Omega}}^{\mathcal{S}\times\mathcal{A}}(d\|d_{\pi_E}),$$

where the equality comes from the definition of Bregman divergence.

Proof of $t(s,a;\pi_d) = \nabla[\bar{\Omega}(d)](s,a)$. For simpler notation, we use matrix-vector notation for the proof when discrete state and action spaces $\mathcal{S} = \{1,...,|\mathcal{S}|\}$ and $\mathcal{A} = \{1,...,|\mathcal{A}|\}$ are considered. For a normalized visitation distribution $d \in \mathcal{D}$, let us define

$$\begin{aligned}
&d_a^s := d(s,a), s\in\mathcal{S}, a\in\mathcal{A}, \\
&\boldsymbol{d}^s := [d_1^s, ..., d_{|\mathcal{A}|}^s]^T \in \mathbb{R}^{\mathcal{A}}, s\in\mathcal{S}, \\
&\boldsymbol{D} := [\boldsymbol{d}^1, ..., \boldsymbol{d}^{|\mathcal{S}|}]^T = \begin{bmatrix} d_1^1 & \cdots & d_{|\mathcal{A}|}^1 \\ \vdots & \ddots & \vdots \\ d_1^{|\mathcal{S}|} & \cdots & d_{|\mathcal{A}|}^{|\mathcal{S}|} \end{bmatrix} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}, \\
&\bar{\boldsymbol{\pi}}(\boldsymbol{x}) := \frac{\boldsymbol{x}}{\boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{x}} = \frac{1}{\sum_{a\in\mathcal{A}} x_a}\left[x_1, ..., x_{|\mathcal{A}|}\right]^T \in \mathbb{R}^{\mathcal{A}}, \boldsymbol{x} := \left[x_1, ..., x_{|\mathcal{A}|}\right]^T \in \mathbb{R}^{\mathcal{A}},
\end{aligned}$$

where $\boldsymbol{1}_{\mathcal{A}} = [1,...,1]^T \in \mathbb{R}^{\mathcal{A}}$ is an $|\mathcal{A}|$-dimensional all-one vector. By using these notations, the original $\bar{\Omega}$ can be rewritten as

$$\bar{\Omega}(\boldsymbol{D}) = \sum_{s,a} d_a^s \Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)) = \sum_{s\in\mathcal{S}} \boldsymbol{1}_{\mathcal{A}}^T \boldsymbol{d}^s \Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)).$$

The gradient of $\bar{\Omega}$ w.r.t. $\boldsymbol{D}$ (using denominator-layout notation) is

$$\nabla_{\boldsymbol{D}}\bar{\Omega}(\boldsymbol{D}) = \left[\frac{\partial\bar{\Omega}(\boldsymbol{D})}{\partial\boldsymbol{d}^1}, ..., \frac{\partial\bar{\Omega}(\boldsymbol{D})}{\partial\boldsymbol{d}^{|\mathcal{S}|}}\right]^T \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}},$$

where each element of $\nabla_{\boldsymbol{D}}\bar{\Omega}(\boldsymbol{D})$ satisfies

$$\begin{aligned}
\frac{\partial\bar{\Omega}(\boldsymbol{D})}{\partial\boldsymbol{d}^s} &= \left[\frac{\partial\bar{\Omega}(\boldsymbol{D})}{\partial d_1^s}, ..., \frac{\partial\bar{\Omega}(\boldsymbol{D})}{\partial d_{|\mathcal{A}|}^s}\right]^T \\
&= \frac{\partial}{\partial\boldsymbol{d}^s}\left\{\sum_{s\in\mathcal{S}} \boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s\Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))\right\} \\
&= \Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))\boldsymbol{1}_{\mathcal{A}} + \boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s\frac{\partial\Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))}{\partial\boldsymbol{d}^s} \\
&= \Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))\boldsymbol{1}_{\mathcal{A}} + \boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s\frac{\partial\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)}{\partial\boldsymbol{d}^s}\frac{\partial\Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))}{\partial\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)}.
\end{aligned} \tag{23}$$

for

$$\begin{aligned}
\frac{\partial\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)}{\partial\boldsymbol{d}^s} &= \left[\frac{\partial\bar{\pi}_1(\boldsymbol{d}^s)}{\partial\boldsymbol{d}^s}, ..., \frac{\partial\bar{\pi}_{|\mathcal{A}|}(\boldsymbol{d}^s)}{\partial\boldsymbol{d}^s}\right], \\
\frac{\partial\bar{\pi}_a(\boldsymbol{d}^s)}{\partial\boldsymbol{d}^s} &= \frac{\partial}{\partial\boldsymbol{d}^s}\left[\frac{d_a^s}{\boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s}\right] = \frac{\partial d_a^s}{\partial\boldsymbol{d}^s}(\boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s)^{-1} + d_a^s\frac{\partial(\boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s)^{-1}}{\partial\boldsymbol{d}^s}.
\end{aligned}$$

Note that each element of $\frac{\partial\bar{\pi}_a(\boldsymbol{d}^s)}{\partial\boldsymbol{d}^s}$ satisfies

$$\begin{aligned}
\frac{\partial\bar{\pi}_a(\boldsymbol{d}^s)}{\partial d_{a'}^s} &= \frac{\partial d_a^s}{\partial d_{a'}^s}(\boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s)^{-1} + d_a^s\frac{\partial(\boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s)^{-1}}{\partial d_{a'}^s} \\
&= \mathbb{I}\{a=a'\}(\boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s)^{-1} - d_a^s(\boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s)^{-2} \\
&= \mathbb{I}\{a=a'\}(\boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s)^{-1} - \bar{\pi}_a(\boldsymbol{d}^s)(\boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s)^{-1} \\
&= (\boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s)^{-1}\left[\mathbb{I}\{a=a'\} - \bar{\pi}_a(\boldsymbol{d}^s)\right],
\end{aligned}$$

and thus,

$$\frac{\partial\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)}{\partial\boldsymbol{d}^s} = (\boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s)^{-1}\left\{\boldsymbol{I}_{\mathcal{A}\times\mathcal{A}} - \boldsymbol{1}_{\mathcal{A}}[\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)]^T\right\}. \tag{24}$$

By substituting Eq.(24) into Eq.(23), we have

$$\begin{aligned}
\frac{\partial\bar{\Omega}(\boldsymbol{D})}{\partial\boldsymbol{d}^s} &= \Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))\boldsymbol{1}_{\mathcal{A}} + \boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s\frac{\partial\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)}{\partial\boldsymbol{d}^s}\frac{\partial\Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))}{\partial\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)} \\
&= \Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))\boldsymbol{1}_{\mathcal{A}} + \boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s\left[(\boldsymbol{1}_{\mathcal{A}}^T\boldsymbol{d}^s)^{-1}\left\{\boldsymbol{I}_{\mathcal{A}\times\mathcal{A}} - \boldsymbol{1}_{\mathcal{A}}[\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)]^T\right\}\right]\frac{\partial\Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))}{\partial\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)} \\
&= \Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))\boldsymbol{1}_{\mathcal{A}} + \left\{\boldsymbol{I}_{\mathcal{A}\times\mathcal{A}} - \boldsymbol{1}_{\mathcal{A}}[\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)]^T\right\}\frac{\partial\Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))}{\partial\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)} \\
&= \Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))\boldsymbol{1}_{\mathcal{A}} + \frac{\partial\Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))}{\partial\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)} - [\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)]^T\frac{\partial\Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))}{\partial\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)}\boldsymbol{1}_{\mathcal{A}} \\
&= \frac{\partial\Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))}{\partial\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)} - [\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)]^T\frac{\partial\Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))}{\partial\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s)}\boldsymbol{1}_{\mathcal{A}} + \Omega(\bar{\boldsymbol{\pi}}(\boldsymbol{d}^s))\boldsymbol{1}_{\mathcal{A}}.
\end{aligned} \tag{25}$$

If we use the function notation, Eq.(25) can be written as

$$\begin{aligned}
\nabla[\bar{\Omega}(d)](s,a) &= \nabla\Omega(\bar{\pi}_d(\cdot|s))(a) - \mathbb{E}_{a'\sim\bar{\pi}_d(\cdot|s)}\left[\nabla\Omega(\bar{\pi}_d(\cdot|s))(a')\right] + \Omega(\bar{\pi}_d(\cdot|s)) \\
&= t(s,a;\bar{\pi}_d)
\end{aligned}$$

for t of Eq.(7) in **Lemma 1**.

E DERIVATION OF BREGMAN-DIVERGENCE-BASED MEASURE IN CONTINUOUS CONTROLS

In Eq.(16), the Bregman divergence in the control task is defined as

$$D_\omega^{\mathcal{A}}(\mathbb{P}_1||\mathbb{P}_2) := \int_{\mathcal{X}} \{\omega(p_1(x)) - \omega(p_2(x)) - \omega'(p_2(x))(p_1(x) - p_2(x))\}\, dx. \tag{26}$$

Note that we consider $\Omega(p) = \int_{\mathcal{X}} \omega(p(x))dx = \int_{\mathcal{X}} [-f_\phi(p(x))]\, dx$ for $f_\phi(x) = x\phi(x)$, which makes Eq.(26) equal to

$$\begin{aligned}
&\int_{\mathcal{X}} \left\{-p_1(x)\phi(p_1(x)) + p_2(x)\phi(p_2(x)) + f'_\phi(p_2(x))(p_1(x) - p_2(x))\right\} dx \\
&= \int_{\mathcal{X}} p_1(x) \left\{f'_\phi(p_2(x)) - \phi(p_1(x))\right\} dx - \int_{\mathcal{X}} p_2(x) \left\{f'_\phi(p_2(x)) - \phi(p_2(x))\right\} dx \\
&= \mathbb{E}_{x\sim p_1}\left[f'_\phi(p_2(x)) - \phi(p_1(x))\right] - \mathbb{E}_{x\sim p_2}\left[f'_\phi(p_2(x)) - \phi(p_2(x))\right].
\end{aligned}$$

Thus, by considering a learning agent's policy $\pi^s = \pi(\cdot|s)$, expert's policy $\pi_E^s = \pi_E(\cdot|s)$, and the objective in Eq.(8) characterized by the Bregman divergence, we can think of the following measure between expert and agent policies:

$$\begin{aligned}
&\mathbb{E}_{s\sim d_\pi}\left[D_\Omega^{\mathcal{A}}(\pi^s||\pi_E^s)\right] \\
&= \mathbb{E}_{s\sim d_\pi}\left[\mathbb{E}_{a\sim\pi^s}\left[f'_\phi(\pi_E^s(a)) - \phi(\pi^s(a))\right] - \mathbb{E}_{a\sim\pi_E^s}\left[f'_\phi(\pi_E^s(a)) - \phi(\pi_E^s(a))\right]\right].
\end{aligned} \tag{27}$$

F TSALLIS ENTROPY AND ASSOCIATED BREGMAN DIVERGENCE AMONG MULTI-VARIATE GAUSSIAN DISTRIBUTIONS

Based on the derivation in Nielsen & Nock (2011), we derive the Tsallis entropy and associated Bremgan divergence as follows. We first consider the distributions in the exponential family

$$\exp\left(\langle\theta, t(x)\rangle - F(\theta) + k(x)\right). \tag{28}$$

Note that for

$$\begin{aligned}
\theta &= \begin{bmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\Sigma^{-1} \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \\
t(x) &= \begin{bmatrix} x \\ xx^T \end{bmatrix}, \\
F(\theta) &= -\frac{1}{4}\theta_1^T\theta_2^{-1}\theta_1 + \frac{1}{2}\log|-\pi\theta_2^{-1}| = \frac{1}{2}\mu^T\Sigma^{-1}\mu + \frac{1}{2}\log(2\pi)^d|\Sigma|, \\
k(x) &= 0,
\end{aligned}$$

we can recover the multi-variate Gaussian distribution (Nielsen & Nock, 2011):

$$\exp(\langle\theta, t(x)\rangle - F(\theta) + k(x)) \tag{29}$$

$$= \exp\left(\mu^T\Sigma^{-1}x - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}xx^T) - \frac{1}{2}\mu^T\Sigma^{-1}\mu - \frac{1}{2}\log(2\pi)^d|\Sigma|\right) \tag{30}$$

$$= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(\mu^T\Sigma^{-1}x - \frac{1}{2}x^T\Sigma^{-1}x - \frac{1}{2}\mu^T\Sigma^{-1}\mu\right) \tag{31}$$

$$= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right). \tag{32}$$

For two distributions with $k(x) = 0$,

$$\pi(x) = \exp(\langle\theta, t(x)\rangle - F(\theta)), \hat{\pi}(x) = \exp(\langle\hat{\theta}, t(x)\rangle - F(\hat{\theta}))$$

that share t, F, and k, it can be shown that

$$\begin{aligned} I(\pi, \hat{\pi}; \alpha, \beta) &= \int \pi(x)^{\alpha} \hat{\pi}(x)^{\beta} dx \\ &= \exp\left(F(\alpha\theta + \beta\hat{\theta}) - \alpha F(\theta) - \beta F(\hat{\theta})\right) \end{aligned}$$

since

$$\begin{aligned} &\int \pi(x)^{\alpha} \hat{\pi}(x)^{\beta} dx \\ =& \int \exp\left(\alpha\langle\theta, t(x)\rangle - \alpha F(\theta) + \beta\langle\hat{\theta}, t(x)\rangle - \beta F(\hat{\theta})\right) dx \\ =& \int \exp\left(\langle\alpha\theta + \beta\hat{\theta}, t(x)\rangle - F(\alpha\theta + \beta\hat{\theta})\right) \exp\left(F(\alpha\theta + \beta\hat{\theta}) - \alpha F(\theta) - \beta F(\hat{\theta})\right) dx \\ =& \exp\left(F(\alpha\theta + \beta\hat{\theta}) - \alpha F(\theta) - \beta F(\hat{\theta})\right) \int \exp\left(\langle\alpha\theta + \beta\hat{\theta}, t(x)\rangle - F(\alpha\theta + \beta\hat{\theta})\right) dx \\ =& \exp\left(F(\alpha\theta + \beta\hat{\theta}) - \alpha F(\theta) - \beta F(\hat{\theta})\right). \end{aligned}$$

F.1 TSALLIS ENTROPY

For $\phi(x) = \frac{k}{q-1}(1 - x^{q-1})$ and $k = 1$, the Tsallis entropy of π can be written as

$$\begin{aligned} \mathcal{T}_q(\pi) := \mathbb{E}_{x\sim\pi}\phi(x) &= \int \pi(x) \frac{1 - \pi(x)^{q-1}}{q-1} dx \\ &= \frac{1 - \int \pi(x)^q dx}{q-1} \\ &= \frac{1}{q-1}\left(1 - I(\pi, \pi; q, 0)\right) = \frac{1 - \exp\left(F(q\theta) - qF(\theta)\right)}{q-1}. \end{aligned}$$

If π is a multivariate Gaussian distribution, we have

$$\begin{aligned} F(q\theta) &= \frac{q}{2}\mu^T\Sigma^{-1}\mu + \frac{1}{2}\log(2\pi)^d|\Sigma| - \frac{1}{2}\log q^d, \\ qF(\theta) &= \frac{q}{2}\mu^T\Sigma^{-1}\mu + \frac{q}{2}\log(2\pi)^d|\Sigma|, \\ F(q\theta) - qF(\theta) &= \frac{1-q}{2}\log(2\pi)^d|\Sigma| - \frac{1}{2}\log q^d \\ &= (1-q)\left\{\frac{d}{2}\log 2\pi + \frac{1}{2}\log|\Sigma| - \frac{d\log q}{2(1-q)}\right\}. \end{aligned}$$

For $\Sigma = \mathrm{diag}\{\sigma_1^2, ..., \sigma_d^2\}$, we have

$$\begin{aligned} F(q\theta) - qF(\theta) &= (1-q)\left\{\frac{d}{2}\log 2\pi + \frac{1}{2}\log|\Sigma| - \frac{d\log q}{2(1-q)}\right\} \\ &= (1-q)\left\{\frac{d}{2}\log 2\pi + \frac{1}{2}\log\prod_{i=1}^{d}\sigma_i^2 - \frac{d\log q}{2(1-q)}\right\} \\ &= (1-q)\sum_{i=1}^{d}\left\{\frac{\log 2\pi}{2} + \log\sigma_i - \frac{\log q}{2(1-q)}\right\}. \end{aligned}$$

F.2 ANALYTICAL FORM OF BASELINE

For $\phi(x) = \frac{k}{q-1}(1 - x^{q-1})$, we have

$$\begin{aligned} f'_\phi(x) &= \frac{k}{q-1}(1 - qx^{q-1}) \\ &= \frac{k}{q-1}(q - qx^{q-1} - (q-1)) \\ &= \frac{qk}{q-1}(1 - x^{q-1}) - k \\ &= q\phi(x) - k. \end{aligned}$$

Therefore, the baseline can be rewritten as

$$\mathbb{E}_{x\sim\pi}[-f'_\phi(x) + \phi(x)] = \mathbb{E}_{x\sim\pi}[k - q\phi(x) + \phi(x)] = (1-q)\mathcal{T}_q(\pi) + k.$$

For a multivariate Gaussian distribution π, the analytical form of $\mathbb{E}_{x\sim\pi}[-f'_\phi(x) + \phi(x)]$ can be derived by using that of Tsallis entropy $\mathcal{T}_q(\pi)$ of π.

F.3 BREGMAN DIVERGENCE WITH TSALLIS ENTROPY REGULARIZATION

In Eq.(27), we consider the following form of the Bregman divergence:

$$\int \pi(x)\{f'_\phi(\hat{\pi}(x)) - \phi(\pi(x))\}dx - \int \hat{\pi}(x)\{f'_\phi(\hat{\pi}(x)) - \phi(\hat{\pi}(x))\}dx.$$

For $\phi(x) = \frac{k}{q-1}(1 - x^{q-1})$, $f'_\phi(x) = \frac{k}{q-1}(1 - qx^{q-1}) = q\phi(x) - k$, and $k = 1$, the above form is equal to

$$\begin{aligned} &\int \pi(x)\left[\frac{1 - q\hat{\pi}(x)^{q-1}}{q-1}\right]dx - \mathcal{T}_q(\pi) - (q-1)\mathcal{T}_q(\hat{\pi}) + 1 \\ &= \frac{1}{q-1} - \frac{q}{q-1}\int \pi(x)\hat{\pi}(x)^{q-1}dx - \mathcal{T}_q(\pi) - (q-1)\mathcal{T}_q(\hat{\pi}) + 1 \\ &= \frac{q}{q-1} - \frac{q}{q-1}\int \pi(x)\hat{\pi}(x)^{q-1}dx - \mathcal{T}_q(\pi) - (q-1)\mathcal{T}_q(\hat{\pi}). \end{aligned}$$

For multivariate Gaussians

$$\begin{aligned} \pi(x) &= \mathcal{N}(x; \mu, \Sigma), \mu = [\nu_1, ..., \nu_d]^T, \Sigma = \mathrm{diag}(\sigma_1^2, ..., \sigma_d^2), \\ \hat{\pi}(x) &= \mathcal{N}(x; \hat{\mu}, \hat{\Sigma}), \hat{\mu} = [\hat{\nu}_1, ..., \hat{\nu}_d]^T, \hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1^2, ..., \hat{\sigma}_d^2), \end{aligned}$$

we have

$$\int \pi(x)\hat{\pi}(x)^{q-1}dx = I(\pi, \hat{\pi}; 1, q-1) = \exp\left(F(\theta') - F(\theta) - (q-1)F(\hat{\theta})\right),$$

where

$$\begin{aligned} \theta &= \begin{bmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\Sigma^{-1} \end{bmatrix}, \\ \hat{\theta} &= \begin{bmatrix} \hat{\Sigma}^{-1}\hat{\mu} \\ -\frac{1}{2}\hat{\Sigma}^{-1} \end{bmatrix}, \\ \theta' = \theta + (q-1)\hat{\theta} &= \begin{bmatrix} \Sigma^{-1}\mu + (q-1)\hat{\Sigma}^{-1}\hat{\mu} \\ -\frac{1}{2}(\Sigma^{-1} + (q-1)\hat{\Sigma}^{-1}) \end{bmatrix} = \begin{bmatrix} \theta'_1 \\ \theta'_2 \end{bmatrix}, \\ \theta'_1 &= \left[\frac{\nu_1}{\sigma_1^2} + (q-1)\frac{\hat{\nu}_1}{\hat{\sigma}_1^2}, ..., \frac{\nu_d}{\sigma_d^2} + (q-1)\frac{\hat{\nu}_d}{\hat{\sigma}_d^2}\right]^T, \\ \theta'_2 &= -\frac{1}{2}\mathrm{diag}\left\{\frac{1}{\sigma_1^2} + (q-1)\frac{1}{\hat{\sigma}_1^2}, ..., \frac{1}{\sigma_d^2} + (q-1)\frac{1}{\hat{\sigma}_d^2}\right\}, \end{aligned}$$

and

$$
\begin{aligned}
F(\theta) &= \frac{1}{2}\mu^T\Sigma^{-1}\mu + \frac{1}{2}\log(2\pi)^d|\Sigma| = \sum_{i=1}^{d}\left\{\frac{\nu_i^2}{2\sigma_i^2} + \frac{\log 2\pi}{2} + \log\sigma_i\right\}, \\
F(\hat{\theta}) &= \frac{1}{2}\hat{\mu}^T\hat{\Sigma}^{-1}\hat{\mu} + \frac{1}{2}\log(2\pi)^d|\hat{\Sigma}| = \sum_{i=1}^{d}\left\{\frac{\hat{\nu}_i^2}{2\hat{\sigma}_i^2} + \frac{\log 2\pi}{2} + \log\hat{\sigma}_i\right\}, \\
F(\theta + (q-1)\hat{\theta}) &= -\frac{1}{4}(\theta_1')^T(\theta_2')^{-1}\theta_1' + \frac{1}{2}\log|-\pi(\theta_2')^{-1}| \\
&= \sum_{i=1}^{d}\left\{\frac{1}{2}\frac{\left(\frac{\nu_i}{\sigma_i^2} + (q-1)\frac{\hat{\nu}_i}{\hat{\sigma}_i^2}\right)^2}{\frac{1}{\sigma_i^2} + (q-1)\frac{1}{\hat{\sigma}_i^2}} + \frac{\log 2\pi}{2} + \log\frac{1}{\frac{1}{\sigma_i^2} + (q-1)\frac{1}{\hat{\sigma}_i^2}}\right\}.
\end{aligned}
$$

G ALGORITHM

Algorithm 1: Regularized Adversarial Inverse Reinforcement Learning (RAIRL)

1: **Input:** A set $\mathcal{D}_E$ of expert demonstration generated by expert's policy π_E, a reward approximator r_θ, a policy π_ψ for neural network parameters θ and ψ
2: **for** each iteration **do**
3: Sample rollout trajectories by using the learners policy π_ψ.
4: Optimize θ with the discriminator D_{r_θ,π_ψ} and the learning objective in Eq.(11).
5: Optimize ψ with Regularized Actor Critic by using r_θ as a reward function.
6: **end for**
7: **Output:** $\pi_\psi \approx \pi_E$, $r_\theta(s,a) \approx t(s,a;\pi_E)$ (a solution of the IRL problem in **Lemma 1**).

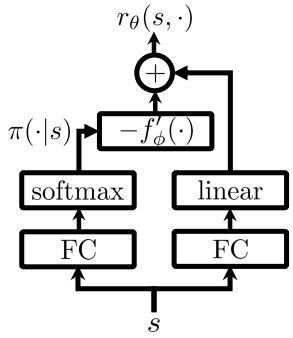
H EXPERIMENT SETTING

H.1 POLICY REGULARIZERS IN EXPERIMENTS

Table 1: Policy regularizers ϕ and their corresponding f_ϕ (Yang et al., 2019).

reg. type.	condition	$\phi(x)$	$f'_\phi(x)$
Shannon	-	$-\log x$	$-\log x - 1$
Tsallis	$k>0, q>1$	$\frac{k}{q-1}(1-x^{q-1})$	$\frac{k}{q-1}(1-qx^{q-1})$
Exp	$k\geq 0, q\geq 1$	$q - x^k q^x$	$q - x^k q^x(k+1+x\log q)$
Cos	$0<\theta\leq\pi/2$	$\cos(\theta x)-\cos(\theta)$	$-\cos(\theta)+\cos(\theta x)-\theta x\sin(\theta x)$
Sin	$0<\theta\leq\pi/2$	$\sin(\theta)-\sin(\theta x)$	$\sin(\theta)-\sin(\theta x)-\theta x\cos(\theta x)$

H.2 DENSITY-BASED MODEL



H.3 EXPERT IN BERMUDA WORLD ENVIRONMENT

We assume a stochastic expert defined by

$$\pi_E(a|s) = \frac{\sum_{t=1}^{3}(d^{(t)})^{-1}\mathbb{I}\{a=\mathrm{Proj}(\theta^{(t)})\}}{\sum_{t=1}^{3}(d^{(t)})^{-1}},$$

$$\theta^{(t)} = \mathrm{arctan2}(\bar{y}^{(t)}-y, \bar{x}^{(t)}-x), d^{(t)} = \|\bar{s}^{(t)}-s\|_2^4+\epsilon, t=1,2,3,$$

for $s=(x,y), \bar{s}^{(1)}=(\bar{x}^{(1)},\bar{y}^{(1)})=(-5,10), \bar{s}^{(2)}=(\bar{x}^{(2)},\bar{y}^{(2)})=(0,10), \bar{s}^{(3)}=(\bar{x}^{(3)},\bar{y}^{(3)})=(5,10)$, $\epsilon=10^{-4}$ and an operator $\mathrm{Proj}(\theta):\mathbb{R}\to\mathcal{A}$ that maps θ to the closest angle in $\mathcal{A}$.

H.4 MuJoCo experiment setting

Instead of directly using MuJoCo environments with $\tanh$-squashed policies proposed in Soft-Actor Critic (Haarnoja et al., 2018), we move $\tanh$ to a part of environment – named *hyperbolized environments* in short – and assume Gaussian policies. Specifically, after an action a is sampled from the policies, we pass $\tanh(a)$ to the environment. We then consider multi-variate Gaussian policy

$$\pi(\cdot|s) = \mathcal{N}\left(\boldsymbol{\mu}(s), \boldsymbol{\Sigma}(s)\right)$$

with $\boldsymbol{\mu}(s) = [\mu_1(s), ..., \mu_d(s)]^T$, $\boldsymbol{\Sigma}(s) = \mathrm{diag}\{(\sigma_1(s))^2, ..., (\sigma_d(s))^2\}$, where

$$-\mathrm{arctanh}(0.99) \leq \mu_i(s) \leq \mathrm{arctanh}(0.99), \log(0.01) \leq \log \sigma_i(s) \leq \log(2)$$

for all $i = 1, ..., d$. Instead of using clipping, we use $\tanh$-activated outputs and scale them to be fit in the above ranges, which empirically improves the performance. Also, instead of using potential-based reward shaping used in AIRL (Fu et al., 2018), we update moving mean of intermediate reward values and update value networks with mean-subtracted rewards to stabilize RL algorithm. Note that this is motivated by **Lemma 2** from which we can guarantee that any constant shift of reward functions does not change optimality.

H.5 Hyperparameters

Table 2, Table 3 and Table 4 list the parameters used in our Bandit, Bermuda World, and MuJoCo experiments, respectively.

Table 2: Hyperparameters for Bandit environments.

Hyper-parameter	Bandit
Batch size	500
Initial exploration steps	$10,000$
Replay size	$500,000$
Target update rate (τ)	0.0005
Learning rate	0.0005
λ	5
q (Tsallis entropy $\mathcal{T}_q^k$)	2.0
k (Tsallis entropy $\mathcal{T}_q^k$)	1.0
Number of trajectories	$1,000$
Reward learning rate	0.0005
Steps per update	50
Total environment steps	$500,000$

TABLE 3: HYPERPARAMETERS FOR BERMUDA WORLD ENVIRONMENT.

HYPER-PARAMETER	BERMUDA WORLD
BATCH SIZE	500
INITIAL EXPLORATION STEPS	$10,000$
REPLAY SIZE	$500,000$
TARGET UPDATE RATE (τ)	0.0005
LEARNING RATE	0.0005
q (TSALLIS ENTROPY $\mathcal{T}_q^k$)	2.0
k (TSALLIS ENTROPY $\mathcal{T}_q^k$)	1.0
NUMBER OF TRAJECTORIES	$1,000$
REWARD LEARNING RATE	0.0005
(FOR EVALUATION) λ	1
(FOR EVALUATION) LEARNING RATE	0.001
(FOR EVALUATION) TARGET UPDATE RATE (τ)	0.0005
STEPS PER UPDATE	50
NUMBER OF STEPS	$500,000$

TABLE 4: HYPERPARAMETERS FOR MUJOCO ENVIRONMENTS.

HYPER-PARAMETER	HOPPER	WALKER2D	HALFCHEETAH	ANT
BATCH SIZE	256	256	256	256
INITIAL EXPLORATION STEPS	$10,000$	$10,000$	$10,000$	$10,000$
REPLAY SIZE	$1,000,000$	$1,000,000$	$1,000,000$	$1,000,000$
TARGET UPDATE RATE (τ)	0.005	0.005	0.005	0.005
LEARNING RATE	0.001	0.001	0.001	0.001
LAMBDA	10^{-4}	10^{-6}	10^{-4}	10^{-6}
k (TSALLIS ENTROPY $\mathcal{T}_q^k$)	1.0	1.0	1.0	1.0
NUMBER OF TRAJECTORIES	100	100	100	100
REWARD LEARNING RATE	0.001	0.001	0.001	0.001
STEPS PER UPDATE	1	1	1	1
NUMBER OF STEPS	$1,000,000$	$1,000,000$	$1,000,000$	$2,000,000$