RIZE: Regularized Imitation Learning via Distributional Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

We propose a novel Inverse Reinforcement Learning (IRL) method that mitigates the rigidity of fixed reward structures and the limited flexibility of implicit reward regularization. Building on the Maximum Entropy IRL framework, our approach incorporates a squared temporal-difference (TD) regularizer with adaptive targets that evolve dynamically during training, thereby imposing adaptive bounds on recovered rewards and promoting robust decision-making. To capture richer return information, we integrate distributional RL into the learning process. Empirically, our method achieves expert-level performance on complex MuJoCo tasks, surpassing baseline methods on the *Humanoid* task with *3 demonstrations*. Extensive experiments and ablation studies further validate the effectiveness of the approach and provide insights into reward dynamics in imitation learning.

1 Introduction

Designing effective reward functions remains a fundamental challenge in Reinforcement Learning (RL). While dense, well-shaped rewards can facilitate learning, they often require laborious task-specific tuning and domain expertise, limiting their scalability (Amodei et al., 2016; Peng et al., 2020). To circumvent these limitations, researchers have explored alternative approaches, including sparse rewards upon task completion (Silver et al., 2016), learning from expert trajectories (Schaal, 1996; Ng & Russell, 2000; Osa et al., 2018), human preference-based reward modeling (Christiano et al., 2017; Lee et al., 2021; Hejna & Sadigh, 2023), and intrinsically motivated RL (Oudeyer et al., 2007; Schmidhuber, 2010; Colas et al., 2022). Among these, Inverse Reinforcement Learning (IRL) (Abbeel & Ng, 2004) offers a compelling alternative by inferring reward functions directly from expert demonstrations, bypassing manual reward engineering. IRL has driven breakthroughs in robotics (Osa et al., 2018), autonomous driving (Knox et al., 2023), and drug discovery (Ai et al., 2024).

A prominent framework in IRL is Maximum Entropy (MaxEnt) IRL (Ziebart, 2010), which underpins many state-of-the-art (SOTA) IL methods. Prior works have combined MaxEnt IRL with adversarial training (Ho & Ermon, 2016; Fu et al., 2018) to minimize divergences between agent and expert distributions. However, these adversarial methods often suffer from instability during training. To address this, recent research has introduced implicit reward regularization, which indirectly represents rewards via Q-values by inverting the Bellman equation. For instance, IQ-Learn (Garg et al., 2021) unifies reward and policy representations using Q-functions with an L_2-norm regularization on rewards, while LSIQ (Al-Hafez et al., 2023) minimizes the chi-squared divergence between expert and mixture distributions, resulting in a squared temporal difference (TD) error objective analogous to SQIL (Reddy et al., 2020). Despite its effectiveness, this method has limitations: LSIQ assigns fixed targets for implicit rewards (e.g., +1 for expert samples and -1 for agent samples), which constrains flexibility by treating all tasks and state-action pairs uniformly, limiting performance and requiring additional gradient steps for convergence.

We propose an extension of implicit reward regularization under the MaxEnt IRL framework, introducing two key advancements. **Adaptive Reward Targets:** We enhance prior TD-error regularization by introducing learnable targets λ^{π_E} and λ^{π} that dynamically adjust during training. These targets replace static constraints with context-sensitive reward alignment, preventing rewards from over-increasing or over-

decreasing. Crucially, our theoretical analysis reveals that these adaptive bounds constrain implicit rewards to well-defined ranges $\left[-\frac{1}{2c} + \min\{\lambda^{\pi_E}, \lambda^{\pi}\}, \frac{1}{2c} + \max\{\lambda^{\pi_E}, \lambda^{\pi}\}\right]$, where c is the regularization coefficient. Since implicit rewards derive from Q-values, this regularization indirectly stabilizes policy training. **Distributional RL Integration:** We incorporate value distributions $Z^{\pi}(s, a)$ (Bellemare et al., 2017) to capture richer uncertainty information in returns, while using their expectations for policy optimization. Though distributional RL has shown success in adversarial IRL (Zhou et al., 2023), it remains unexplored in non-adversarial settings. Our work bridges this gap, demonstrating its efficacy in MaxEnt IRL while preserving theoretical guarantees. Unifying these advances, our framework outperforms online IL baselines on MuJoCo benchmarks (Todorov et al., 2012). Notably, our approach proves vital for complex tasks like *Humanoid*, where both adaptive rewards and distributional learning are essential.

Our contributions are threefold. First, we introduce adaptive targets for implicit reward regularization, enabling dynamic reward bounds that enhance stability and alignment during training. Second, we integrate value distributions into implicit reward frameworks, capturing richer uncertainty information while preserving theoretical consistency. Third, we empirically validate our approach through extensive MuJoCo experiments, demonstrating superior performance in complex environments.

2 Related Work

Imitation learning (IL) and inverse reinforcement learning (IRL) (Watson et al., 2023) are foundational paradigms for training agents to mimic expert behavior from demonstrations. Behavioral Cloning (BC) (Pomerleau, 1991), the simplest IL approach, treats imitation as a supervised learning problem by directly mapping states to expert actions. While computationally efficient, BC is prone to compounding errors (Ross & Bagnell, 2011) due to covariate shift during deployment. The Maximum Entropy IRL framework (Ziebart, 2010) addresses this limitation by probabilistically modeling expert behavior as reward maximization under an entropy regularization constraint, establishing a theoretical foundation for modern IRL methods.

The advent of adversarial training marked a pivotal shift in IL methodologies. Ho & Ermon (2016) introduced Generative Adversarial Imitation Learning (GAIL), which formulates imitation learning as a generative adversarial game (Goodfellow et al., 2014) where an agent learns a policy indistinguishable from the expert's by minimizing the Jensen-Shannon divergence between their state-action distributions. This framework was generalized by Ghasemipour et al. (2019) in f-GAIL, which replaces the Jensen-Shannon divergence with arbitrary f-divergences to broaden applicability. Concurrently, Kostrikov et al. (2019) proposed Discriminator Actor-Critic (DAC), improving sample efficiency via off-policy updates and terminal-state reward modeling while mitigating reward bias.

Recent advances have shifted toward methods that bypass explicit reward function estimation. Kostrikov et al. (2020) introduced ValueDICE, an offline IL method that leverages an inverse Bellman operator to avoid adversarial optimization. Similarly, Garg et al. (2021) developed IQ-Learn, which circumvents the challenges of MaxEnt IRL by optimizing implicit rewards derived directly from expert Q-values. A parallel research direction simplifies reward engineering by assigning fixed rewards to expert and agent samples. Reddy et al. (2020) pioneered this approach with Soft Q Imitation Learning (SQIL), which assigns binary rewards to transitions from expert and agent trajectories. Most recently, Al-Hafez et al. (2023) proposed Least Squares Inverse Q-Learning (LSIQ), enhancing regularization by minimizing chi-squared divergence between expert and mixture distributions while explicitly managing absorbing states through critic regularization. Our work builds on these foundations by integrating distributional RL (Dabney et al., 2018a) and introducing a squared TD regularizer with adaptive targets.

3 Background

3.1 Preliminary

We consider a Markov Decision Process (MDP) (Puterman, 2014) to model policy learning in Reinforcement Learning (RL). The MDP framework is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, p_0, P, R, \gamma \rangle$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action space, p_0 the initial state distribution, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ the transition kernel

with $P(\cdot \mid s, a)$ specifying the likelihood of transitioning from state s given action a, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function, and $\gamma \in [0, 1]$ the discount factor which tempers future rewards. A stationary policy $\pi \in \Pi$ is characterized as a mapping from states $s \in \mathcal{S}$ to distributions over actions $a \in \mathcal{A}$. The primary objective in RL (Sutton & Barto, 2018) is to maximize the expected sum of discounted rewards, expressed as $\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$. Furthermore, the occupancy measure $\rho_\pi(s, a)$ for a policy $\pi \in \Pi$ is given by $(1 - \gamma)\pi(a \mid s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \pi)$. The corresponding measure for an expert policy, π_E, is similarly denoted by ρ_E. In Imitation Learning (IL), the expert policy π_E is typically unknown, and only a finite set of expert demonstrations is available, rather than explicit reward feedback from the environment.

3.2 Distributional Reinforcement Learning

Maximum Entropy (MaxEnt) RL (Haarnoja et al., 2018) focuses on addressing the stochastic nature of action selection by maximizing the entropy of the policy, while Distributional RL (Bellemare et al., 2017) emphasizes capturing the inherent randomness in returns. Combining these perspectives, the distributional soft value function $Z : \mathcal{S} \times \mathcal{A} \to \mathcal{Z}$ (Ma et al., 2020) for a policy $\pi \in \Pi$ encapsulates uncertainty in both rewards and actions, with $\mathcal{Z}$ representing the space of action-value distributions. It is formally defined as:

$$Z(s, a) = \sum_{t=0}^{\infty} \gamma^t [R(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot \mid s_t))], \tag{1}$$

where $\mathcal{H}(\pi) = \mathbb{E}_\pi[-\log \pi(a \mid s)]$ denotes the entropy of the policy, and $\alpha > 0$ balances entropy with reward.

The distributional soft Bellman operator $\mathcal{B}_D^\pi : \mathcal{Z} \to \mathcal{Z}$ for a given policy π is introduced as $(\mathcal{B}_D^\pi Z)(s, a) \overset{D}{=} R(s, a) + \gamma[Z(s', a') - \alpha \log \pi(a' \mid s')]$, where $s' \sim P(\cdot \mid s, a)$, $a' \sim \pi(\cdot \mid s')$, and $\overset{D}{=}$ signifies equality in distribution. Notably, this operator exhibits contraction properties under the p-Wasserstein metric, ensuring convergence to the optimal distributional value function.

A practical approach to approximating the value distribution Z involves modeling its quantile function $F_Z^{-1}(\tau)$, evaluated at specific quantile levels $\tau \in [0, 1]$ (Dabney et al., 2018b). The quantile function is defined as $F_Z^{-1}(\tau) = \inf\{z \in \mathbb{R} : \tau \leq F_Z(z)\}$, where $F_Z(z) = \mathbb{P}(Z \leq z)$ is the cumulative distribution function of Z. For simplicity, we denote the quantile-based representation as $Z_\tau(s, a) := F_Z^{-1}(\tau)$. To discretize this representation, we define a sequence of quantile levels, denoted as $\{\tau_i\}_{i=0,\ldots,N-1}$, where $0 = \tau_0 < \ldots < \tau_{N-1} = 1$. These quantiles partition the unit interval into N fractions. Sampling uniformly from this interval $\tau \sim U(0, 1)$, we obtain an approximation of the value distribution $Z_\tau(s, a)$, which provides quantile values at these specified levels.

3.3 Inverse Reinforcement Learning

Given expert trajectory data, Maximum Entropy (MaxEnt) Inverse RL (Ziebart, 2010) aims to infer a reward function $R(s, a)$ from the family $\mathcal{R} = \mathbb{R}^{S \times A}$. Instead of assuming a deterministic expert policy, this method optimizes for stochastic policies $\pi \in \Pi$ that maximize R while matching expert behavior. GAIL (Ho & Ermon, 2016) extends this framework by introducing a convex reward regularizer $\psi : \mathbb{R}^{S \times A} \to \bar{\mathbb{R}}$, leading to the adversarial objective:

$$\max_{R \in \mathcal{R}} \min_{\pi \in \Pi} L(\pi, R) = \mathbb{E}_{\rho_E}[R(s, a)] - \mathbb{E}_{\rho_\pi}[R(s, a)] - \mathcal{H}(\pi) - \psi(R)\,. \tag{2}$$

IQ-Learn (Garg et al., 2021) departs from adversarial training by implicitly representing rewards through Q-functions $Q \in \Omega$ (Piot et al., 2014). It leverages the inverse soft Bellman operator $\mathcal{T}^\pi$, defined as:

$$(\mathcal{T}^\pi Q)(s, a) = Q(s, a) - \gamma \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi(\cdot|s')}[Q(s', a') - \alpha \log \pi(a'|s')]. \tag{3}$$

For a fixed policy π, $\mathcal{T}^\pi$ is bijective, ensuring a one-to-one correspondence between Q-values and rewards: $\mathcal{T}^\pi Q = R$ and $Q = (\mathcal{T}^\pi)^{-1} R$. This allows reframing the MaxEnt IRL objective (2) in Q-policy space as $\max_{Q \in \Omega} \min_{\pi \in \Pi} \mathcal{J}(\pi, Q)$. IQ-Learn simplifies the problem by defining the implicit reward $R_Q(s, a) =$

$\mathcal{T}^\pi Q(s,a)$ and applying an L2 regularizer $\psi(R_Q)$. The final objective becomes:

$$\begin{aligned}\max_{Q\in\Omega}\min_{\pi\in\Pi}\mathcal{J}(\pi,Q) = \; & \mathbb{E}_{\rho_E}[R_Q(s,a)] - \mathbb{E}_{\rho_\pi}[R_Q(s,a)] - \alpha\mathcal{H}(\pi) \\ & - c\Big[\mathbb{E}_{\rho_E}[R_Q(s,a)^2] + \mathbb{E}_{\rho_\pi}[R_Q(s,a)^2]\Big]\,.\end{aligned} \tag{4}$$

4 Methodology

This section introduces a framework that integrates Distributional Reinforcement Learning with Inverse RL. We first show how value distributions can replace point-estimate critics. We then propose an adaptive regularization technique for implicit rewards and analyze its properties. Finally, we derive RIZE, which combines distributional critics with bounded-reward imitation learning.

4.1 Distributional Value Integration

Our approach departs from traditional imitation learning by explicitly using value distributions as critics within an actor-critic framework (Zhou et al., 2023). We argue that learning the soft value distribution $Z(s,a)$ in Equation (1), rather than relying solely on point estimates like $Q(s,a)$, enhances decision-making by capturing uncertainty in complex environments. This view is consistent with recent neuroscience findings suggesting that decision-making in the prefrontal cortex relies on learning distributions over outcomes rather than only their expectations (Muller et al., 2024).

Moreover, access to the full return distribution enables the computation of statistical moments—most notably the expectation—which we optimize both policy and critic using the expectation of Z (Bellemare et al., 2017; Dabney et al., 2018b;a), yielding a more robust learning signal while remaining compatible with IQ-Learn.

We compute Q as the expectation of the soft value distribution:

$$Q(s,a) = \sum_{i=0}^{N-1}(\tau_{i+1}-\tau_i)\,Z_{\tau_i}(s,a)\,, \tag{5}$$

where $\{\tau_i\}$ are quantile fractions and $Z_{\tau_i}(s,a)$ denotes the corresponding quantile values (see Lemma A.1).

4.2 Implicit Reward Regularization

In this section, we propose a regularizer for inverse RL that refines existing implicit reward formulations (Garg et al., 2021). The implicit reward is defined as:

$$R_Q(s,a) = Q(s,a) - \gamma\mathbb{E}_{P,\pi}\left[Q(s',a') - \alpha\log\pi(a'|s')\right]. \tag{6}$$

Previous works typically regularize implicit rewards either using L_2-norms (Garg et al., 2021) or by treating them as squared-TD errors between rewards and fixed targets (Reddy et al., 2020; Al-Hafez et al., 2023). While we adopt a similar squared-TD setting, we introduce adaptive targets λ^{π_E} (for the expert π_E) and λ^π (for the imitation policy π) to construct our convex regularizer $\Gamma:\mathbb{R}^{S\times A}\to\bar{\mathbb{R}}$:

$$\Gamma(R_Q,\lambda) = \mathbb{E}_{\rho_E}\left[(R_Q(s,a)-\lambda^{\pi_E})^2\right] + \mathbb{E}_{\rho_\pi}\left[(R_Q(s,a)-\lambda^{\pi})^2\right]. \tag{7}$$

These targets self-update through a feedback loop where reward estimates continuously adapt to match moving targets:

$$\min_{\lambda^{\pi_E}}\mathbb{E}_{\rho_E}\left[\left(R_Q(s,a)-\lambda^{\pi_E}\right)^2\right], \qquad \min_{\lambda^{\pi}}\mathbb{E}_{\rho_\pi}\left[\left(R_Q(s,a)-\lambda^{\pi}\right)^2\right]. \tag{8}$$

Substituting $\Gamma(R_Q,\lambda)$ for the L_2 term in Equation (4) and using Equation (5) to compute Q, we obtain:

$$\begin{aligned}\mathcal{L}(\pi,Q) = \; & \mathbb{E}_{\rho_E}[R_Q(s,a)] - \mathbb{E}_{\rho_\pi}[R_Q(s,a)] - \alpha\mathcal{H}(\pi) \\ & - c\Big[\mathbb{E}_{\rho_E}\left[(R_Q(s,a)-\lambda^{\pi_E})^2\right] + \mathbb{E}_{\rho_\pi}\left[(R_Q(s,a)-\lambda^{\pi})^2\right]\Big]\,,\end{aligned} \tag{9}$$

where c is the regularization coefficient.

As in IQ-Learn and LSIQ, our method seeks behavior indistinguishable from expert demonstrations in a non-adversarial implicit-reward setting. Prior work analyzes optimal implicit rewards, drawing on Max–Min analyses from GANs (Al-Hafez et al., 2023; Goodfellow et al., 2014). Because we bind rewards to adaptive targets via $\Gamma(R_Q, \lambda)$ (Equation (7)), we carry out an analogous analysis, stated below.

Proposition 4.1. *Let $R_Q(s,a) = (\mathcal{T}^\pi Q)(s,a)$ denote the implicit reward derived from point-estimate Q-values, where $Q(s,a) = \mathbb{E}[Z(s,a)]$. Let $\rho_E(s,a)$ and $\rho_\pi(s,a)$ denote occupancy measures under π_E and π, respectively. For fixed π, the optimal TD-regularized reward satisfies:*

$$R_Q^*(s,a) = \frac{\rho_E(s,a) - \rho_\pi(s,a)}{(2c)\,(\rho_E(s,a) + \rho_\pi(s,a))} + \frac{\rho_E(s,a)\lambda^{\pi_E} + \rho_\pi(s,a)\lambda^\pi}{\rho_E(s,a) + \rho_\pi(s,a)}\,. \tag{10}$$

Proof. Differentiating $\mathcal{L}(\pi, Q)$ in Equation (9) with respect to $R_Q(s,a)$ (holding π fixed) and setting the result to zero yields

$$0 = \rho_E - \rho_\pi - 2c\big[\rho_E(R_Q - \lambda^{\pi_E}) + \rho_\pi(R_Q - \lambda^\pi)\big]\,, \tag{11}$$

from which Equation (10) follows. □

As established in Corollary A.2, this optimal reward satisfies:

$$R_Q^*(s,a) \in \left[-\frac{1}{2c} + \lambda_{\min},\ \frac{1}{2c} + \lambda_{\max}\right]\,, \tag{12}$$

where $\lambda_{\min} := \min\{\lambda^{\pi_E}, \lambda^\pi\}$ and $\lambda_{\max} := \max\{\lambda^{\pi_E}, \lambda^\pi\}$. The coefficient c and targets bound rewards within a well-defined range. In practice, we use $\lambda \in [0, 10]$ and $c \in [0.1, 0.5]$, which we found to promote stable training (see Appendix B.3).

Moreover, when the occupancy measures match, the optimal reward simplifies significantly, as stated in Corollary A.3:

$$R_Q^*(s,a) = \lambda^{\pi_E} = \lambda^\pi\,. \tag{13}$$

Recalling that we represent rewards through Q-values, the boundedness of the optimal reward ensures that critic updates remain constrained and—because the policy directly depends on these critic values—promotes stable policy optimization. However, as previously noted (Al-Hafez et al., 2023; Viano et al., 2022), the convergence guarantee originally stated for IQ-Learn (Garg et al., 2021) does *not* extend to the χ^2-regularizer used in practice. Consequently, a formal proof for the resulting alternating SAC updates remains an open question and is left for future work.

4.3 Practical Algorithm

We now introduce **RIZE**: Regularized Imitation Learning via Distributional Reinforcement Learning (Algorithm 1). We approximate Z and π with neural networks, optimizing via $Q = \mathbb{E}[Z]$. Following Distributional SAC (Ma et al., 2020), we employ target policies for next-action sampling and a double-critic architecture with target networks for stability. Lower learning rates proved critical for robust policy updates, with a four-layer MLP policy network necessary for complex tasks.

Targets λ^{π_E} and λ^π are optimized to match expected rewards. We initialize λ^{π_E} higher than λ^π and update λ^π with lower learning rates due to its sensitivity.

Algorithm 1 RIZE

Initialize Z_ϕ, π_θ, λ^{π_E}, and λ^π
for step t in $\{1, \ldots, N\}$ **do**
  Calculate $Q(s,a) = \mathbb{E}[Z_\phi(s,a)]$ using Eq. 5
  Update Z_ϕ using Eq. 9
  $\phi_{t+1} \leftarrow \phi_t - \beta_Z \nabla_\phi[-\mathcal{L}(\phi)]$
  Update π_θ (like SAC)
  $\theta_{t+1} \leftarrow \theta_t + \beta_\pi \nabla_\theta \mathbb{E}_{\substack{s\sim\mathcal{D},\\ a\sim\pi_\theta(\cdot|s)}}\big[\min_{k=1,2} Q_k(s,a) - \alpha \log \pi_\theta(a|s)\big]$
  Update λ^π and λ^{π_E} using Eq. 8
  $\lambda^\pi_{t+1} \leftarrow \lambda^\pi_t - \beta_{\lambda^\pi} \nabla_{\lambda^\pi} \Gamma(R_Q, \lambda)$
  $\lambda^{\pi_E}_{t+1} \leftarrow \lambda^{\pi_E}_t - \beta_{\lambda^{\pi_E}} \nabla_{\lambda^{\pi_E}} \Gamma(R_Q, \lambda)$
end for

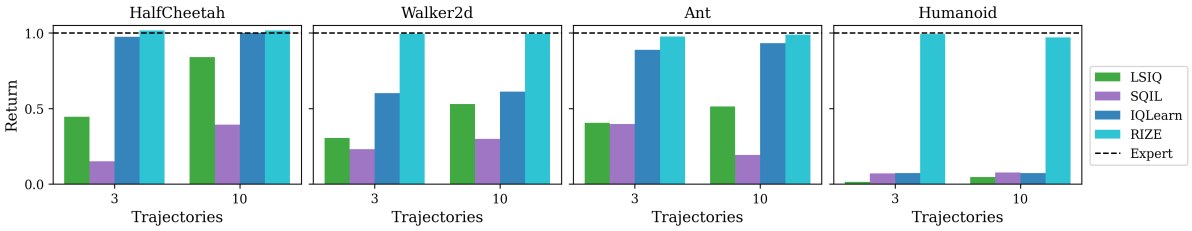


Figure 1: Normalized return of RIZE versus online imitation learning baselines on Gym MuJoCo tasks. We depict the sorted top 25% episodic returns across five seeds to evaluate convergence to expert-level performance, using three and ten expert trajectories.

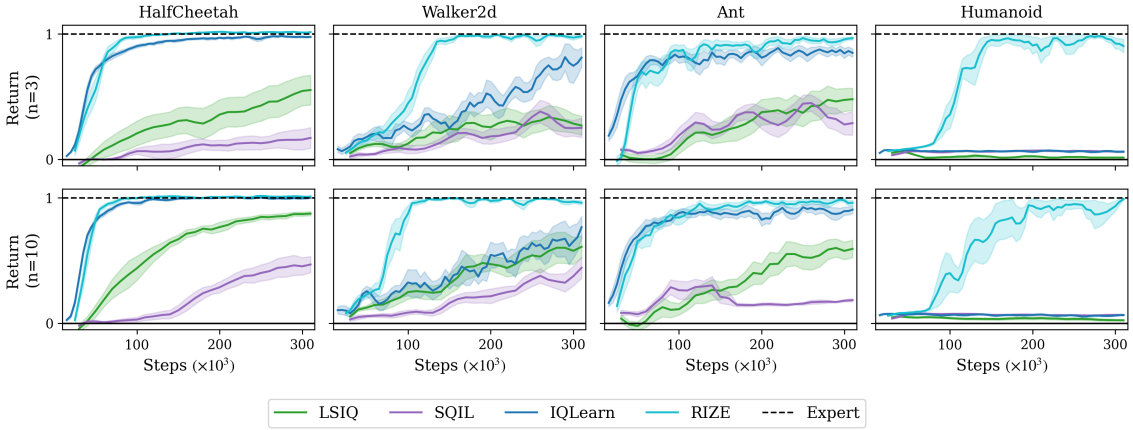


Figure 2: Normalized return of RIZE vs. online imitation learning baselines on Gym MuJoCo tasks. n denotes the number of expert trajectories.

5 Experiments

We evaluate our algorithm on four MuJoCo (Todorov et al., 2012) benchmarks (HalfCheetah-v2, Walker2d-v2, Ant-v2, Humanoid-v2) against state-of-the-art imitation learning methods: IQ-Learn (Garg et al., 2021), LSIQ (Al-Hafez et al., 2023), and SQIL (Reddy et al., 2020). All experiments are conducted with five seeds for statistical significance (Henderson et al., 2018). We assess each method using three and ten expert trajectories. For ten trajectories, we retain baseline hyperparameters; for three trajectories (not reported in prior works), we adapt configurations from single-demonstration settings. IQ-Learn and LSIQ use their official implementations, while SQIL is evaluated using the LSIQ codebase. Results report mean performance across five seeds, with half a standard deviation to indicate uncertainty, and lines are smoothed for better readability. We normalize episode returns based on expert performance. For details, expanded visualizations, and hyperparameter tuning, see Appendix B.

5.1 Main Results

Our method outperforms LSIQ and SQIL across tasks, with IQ-Learn as the only competitive baseline. Notably, in the most complex environment, Humanoid-v2, our approach is the sole method achieving expert-level performance, while all baselines fail (see Figures 1, 2). This demonstrates our algorithm's scalability to high-dimensional control problems. Additionally, our method shows superior sample efficiency, requiring fewer gradient steps to match expert performance compared to SOTA algorithms. These results highlight the robustness and efficiency of our approach in tackling complex tasks.

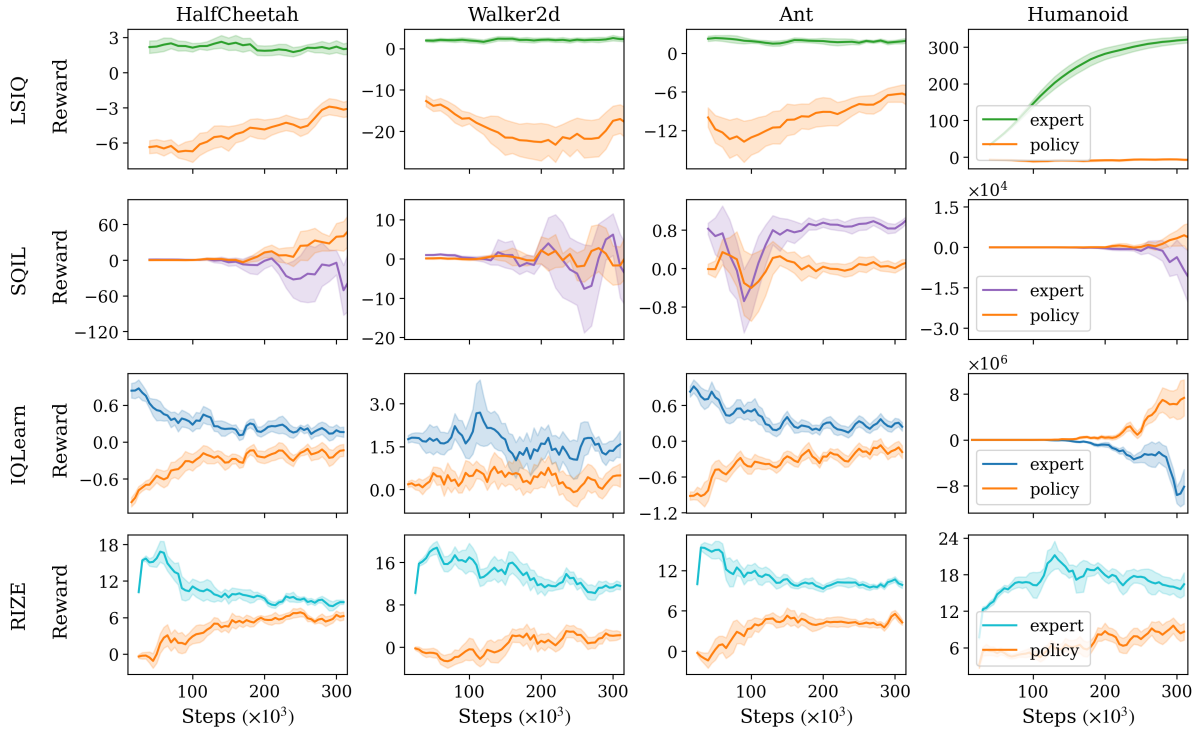


Figure 3: Implicit reward curves for expert and policy samples on MuJoCo tasks. We evaluate on ten expert demonstrations.

5.2 Recovered Reward

Here, we compare the reward behavior of RIZE against the baselines. As shown in Figure 3, our approach effectively bounds the rewards for both expert and policy samples across environments. The reward regularizer $\Gamma(R_Q, \lambda)$ keeps the recovered rewards close to adaptive targets, with the optimization of λ^{π_E} and λ^{π} via Equation (8) enforcing an upper bound on expert rewards and a lower bound on policy rewards. With sufficient expert trajectories, rewards for both expert and policy samples stabilize around consistent values. In contrast, baseline methods exhibit unbounded reward growth in the Humanoid task, and SQIL fails to constrain rewards even in the simpler HalfCheetah environment.

5.3 Role of Distributional Value Function

We examine the evolution of critics in RIZE compared to the baselines. Unlike the baselines, which train Q-networks, RIZE employs distributional value functions and uses their expectation $\mathbb{E}[Z(s, a)]$ for loss optimization. Since all methods update policies by maximizing state-action values, unbounded values can significantly undermine robustness. As illustrated in Figure 4, RIZE and LSIQ maintain bounded state-action values in the challenging Humanoid environment, whereas SQIL and IQ-Learn do not. Nevertheless, the boundedness of critics alone does not guarantee expert-like policy performance, as evidenced by LSIQ's results in Figure 2.

Moreover, to assess the significance of quantile-based value networks $Z(s, a)$, we train RIZE with point-estimate critics $Q(s, a)$ and compare the results with the original implementation. As shown in Figure 5, distributional value functions consistently outperform Q-networks across all tasks, exhibiting lower variance and greater sample efficiency. Notably, in the Humanoid environment, Q-values fail to replicate expert demonstrations, whereas quantile-based critics maintain expert-level performance.

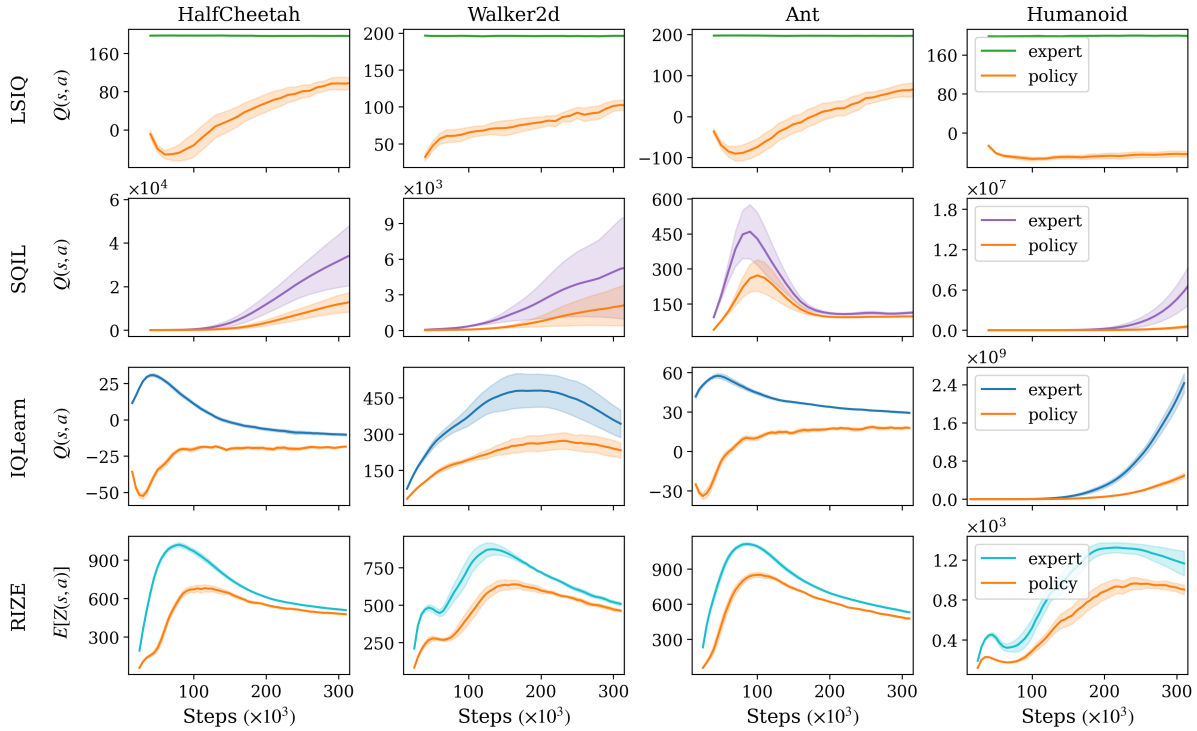


Figure 4: State-action value curves for expert and policy samples on MuJoCo tasks. Ten trajectories are used throughout the analysis.

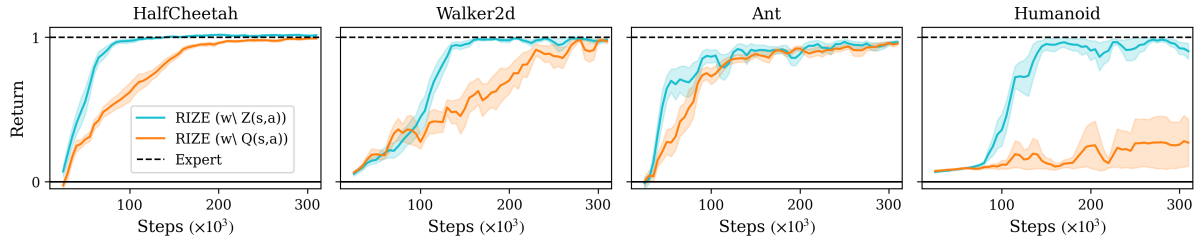


Figure 5: Normalized return of RIZE on Gym MuJoCo tasks utilizing value distribution $Z(s, a)$ or classic $Q(s, a)$ as the critic. We use three expert trajectories for all tasks.

5.4 Ablation: Regularization Strategies

We investigate the effect of squared-TD error regularization on the Ant-v2 task using three distinct formulations: *Expert-focused TD:* $\mathbb{E}_{\rho_E}[R_Q - \lambda^{\pi_E}]^2 + \mathbb{E}_{\rho_\pi}[R_Q]^2$, *Policy-focused TD:* $\mathbb{E}_{\rho_E}[R_Q]^2 + \mathbb{E}_{\rho_\pi}[R_Q - \lambda^{\pi}]^2$, *Baseline L2:* $\mathbb{E}_{\rho_E}[R_Q]^2 + \mathbb{E}_{\rho_\pi}[R_Q]^2$.

Our results in Figure 6 demonstrate that applying squared-TD regularization to both expert and policy samples improves robustness. While the expert-focused TD formulation outperforms the policy-focused variant, neither consistently achieves expert-level behavior. Over time, both exhibit gradual performance degradation, underscoring the challenges of maintaining stable imitation. This highlights the need for more refined regularization strategies to ensure consistent and reliable learning in imitation learning.

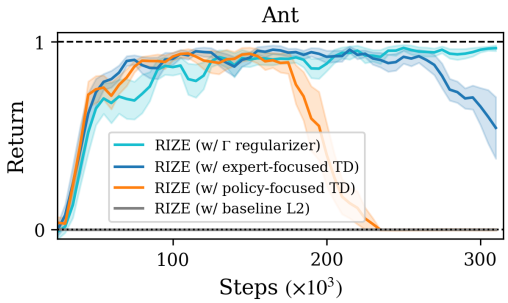


Figure 6: Ablation study on the effect of various regularization strategies, evaluated using three demonstrations.

6 Conclusion

We propose a novel IRL framework that overcomes the limitations of fixed reward mechanisms through dynamic reward adaptation and context-sensitive regularization. Our approach ensures bounded implicit rewards and stable value function updates, leading to robust policy optimization. By integrating distributional RL with implicit reward learning, we capture richer return dynamics while preserving theoretical guarantees. Empirical results on MuJoCo benchmarks show expert-like proficiency on the *Humanoid* task with only *three demonstrations*. Ablation studies confirm the critical role of our regularization mechanism. This work unifies implicit reward regularization with distributional return representations, offering a scalable and sample-efficient solution for complex decision-making. Future directions include extending this framework to offline IL and risk-sensitive robotics, where adaptive rewards and uncertainty-aware distributions can further improve robustness and generalization.

References

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning*, 2004.

Cheng Ai, Hong Yang, Xinyu Liu, Ruijie Dong, Ying Ding, and Fei Guo. Mtmol-gpt: De novo multi-target molecular generation with transformer-based generative adversarial imitation learning. *PLoS Computational Biology*, 20(6):e1012229, 2024. doi: 10.1371/journal.pcbi.1012229. URL https://doi.org/10.1371/journal.pcbi.1012229.

Firas Al-Hafez, Davide Tateo, Oleg Arenz, Guoping Zhao, and Jan Peters. LS-IQ: Implicit reward regularization for inverse reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=o3Q4m8jg4BR.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016. URL http://arxiv.org/abs/1606.06565.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:966543.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: A short survey. *Journal of Artificial Intelligence Research*, 74, July 2022. doi: 10.1613/jair.1.13554. URL https://hal.science/hal-03901771.

Will Dabney, Georg Ostrovski, David Silver, and Remi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 1096–1105. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Remi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018b. ISBN 978-1-57735-800-8.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.

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

Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, 2019.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, 2018.

Joey Hejna and Dorsa Sadigh. Inverse preference learning: Preference-based RL without a reward function. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=gAP52Z2dar`.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

W. Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis)design for autonomous driving. *Artif. Intell.*, 316(C), March 2023. ISSN 0004-3702. doi: 10.1016/j.artint.2022.103829. URL `https://doi.org/10.1016/j.artint.2022.103829`.

Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. In *International Conference on Learning Representations*, 2019.

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. In *International Conference on Learning Representations*, 2020.

Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In *Lee, Kimin and Smith, Laura and Abbeel, Pieter*, 2021.

Xiaoteng Ma, Qiyuan Zhang, Li Xia, Zhengyuan Zhou, Jun Yang, and Qianchuan Zhao. Dsac: Distributional soft actor critic for risk-sensitive reinforcement learning. *arXiv*, 2020.

Timothy H. Muller, James L. Butler, Sebastijan Veselic, Bruno Miranda, Joni D. Wallis, Peter Dayan, Timothy E. J. Behrens, Zeb Kurth-Nelson, and Steven W. Kennerley. Distributional reinforcement learning in prefrontal cortex. *Nature Neuroscience*, 27(3):403–408, mar 2024. ISSN 1546-1726. doi: 10.1038/s41593-023-01535-w. URL `https://doi.org/10.1038/s41593-023-01535-w`.

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, 2000.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 2018.

Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286, 2007. doi: 10.1109/TEVC.2006.890271.

Xue Bin Peng, Erwin Coumans, Tingnan Zhang, Tsang-Wei Edward Lee, Jie Tan, and Sergey Levine. Learning agile robotic locomotion skills by imitating animals. *CoRR*, abs/2004.00784, 2020. URL https://arxiv.org/abs/2004.00784.

B. Piot, M. Geist, and O. Pietquin. Boosted and Reward-regularized Classification for Apprenticeship Learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2014.

D. A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Siddharth Reddy, Anca D Dragan, and Sergey Levine. Sqil: Imitation learning via reinforcement learning with sparse rewards. In *International Conference on Learning Representations*, 2020.

Stéphane Ross and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Stefan Schaal. Learning from demonstration. In M.C. Mozer, M. Jordan, and T. Petsche (eds.), *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL https://proceedings.neurips.cc/paper_files/paper/1996/file/68d13cf26c4b4f4f932e3eff990093ba-Paper.pdf.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010. doi: 10.1109/TAMD.2010.2056368.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. doi: 10.1038/nature16961. URL https://doi.org/10.1038/nature16961.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, Massachusetts; London, England, second edition, 2018.

E. Todorov, T. Erez, and Y. Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012.

Luca Viano, Angeliki Kamoutsi, Gergely Neu, Igor Krawczuk, and Volkan Cevher. Proximal point imitation learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.

Joe Watson, Sandy Huang, and Nicolas Heess. Coherent soft imitation learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=kCCD8d2aEu.

Y. Zhou, M. Lu, X. Liu, Z. Che, Z. Xu, J. Tang, and Y Zhang. Distributional generative adversarial imitation learning with reproducing kernel generalization. *Neural Networks*, 165, 2023.

Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. PhD thesis, University of Washington, 2010.

A Supporting Proofs

A.1 Expectation of the Value Distribution

Lemma A.1. *The expectation of the distributional return satisfies*

$$\mathbb{E}[Z(s,a)] = Q(s,a).$$

Proof. From the soft value distribution definition

$$Z(s,a) = \sum_{t=0}^{\infty} \gamma^t \left[R(s_t,a_t) + \alpha\mathcal{H}(\pi(\cdot \mid s_t))\right] , \tag{14}$$

taking expectations yields:

$$\begin{aligned}\mathbb{E}[Z(s,a)] &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t,a_t) + \alpha\mathcal{H}(\pi(\cdot \mid s_t))\right)\right] \\ &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[R(s_t,a_t) + \alpha\mathcal{H}(\pi(\cdot \mid s_t))\right] \\ &= Q(s,a) \quad \text{(by soft Q-value definition).}\end{aligned}$$

□

A.2 Bounds on the Optimal Implicit Reward

Corollary A.2. *The optimal implicit reward satisfies:*

$$R_Q^*(s,a) \in \left[-\frac{1}{2c} + \lambda_{\min},\ \frac{1}{2c} + \lambda_{\max}\right] , \tag{15}$$

where $\lambda_{\min} := \min\{\lambda^{\pi_E}, \lambda^{\pi}\}$ *and* $\lambda_{\max} := \max\{\lambda^{\pi_E}, \lambda^{\pi}\}$.

Proof. From Proposition 4.1:

$$R_Q^*(s,a) = \underbrace{\frac{\rho_E - \rho_\pi}{2c(\rho_E + \rho_\pi)}}_{\text{Term 1}} + \underbrace{\frac{\rho_E \lambda^{\pi_E} + \rho_\pi \lambda^{\pi}}{\rho_E + \rho_\pi}}_{\text{Term 2}} . \tag{16}$$

Term 1 lies in $\left[-\frac{1}{2c}, \frac{1}{2c}\right]$ since $\rho_E, \rho_\pi \geq 0$ (achieved when either $\rho_\pi \to 0$ or $\rho_E \to 0$). Term 2 is a convex combination of λ^{π_E} and λ^{π}, thus in $[\lambda_{\min},\ \lambda_{\max}]$. Combining intervals yields the result. □

A.3 Convergence of the Optimal Reward

Corollary A.3. *For* $\rho_\pi = \rho_E$*, the optimal reward satisfies* $R_Q^*(s,a) = \lambda^{\pi_E} = \lambda^{\pi}$.

Proof. When $\rho_\pi = \rho_E$, Proposition 4.1 simplifies to

$$R_Q^*(s,a) = \frac{\lambda^{\pi_E} + \lambda^{\pi}}{2} . \tag{17}$$

Substituting this expression into the loss functions for λ^{π_E} and λ^{π} in Equation (8) yields

$$\min_{\lambda^{\pi_E}} \mathbb{E}_{\rho_E}\left[\frac{1}{4}\left(\lambda^{\pi} - \lambda^{\pi_E}\right)^2\right] , \qquad \min_{\lambda^{\pi}} \mathbb{E}_{\rho_\pi}\left[\frac{1}{4}\left(\lambda^{\pi_E} - \lambda^{\pi}\right)^2\right] . \tag{18}$$

Differentiating each objective with respect to its corresponding target and setting the derivatives to zero gives $\lambda^{\pi} = \lambda^{\pi_E}$. This result implies that the optimal targets coincide at convergence. Therefore, the optimal reward reduces to

$$R_Q^*(s,a) = \lambda^{\pi_E} = \lambda^{\pi} . \tag{19}$$

□

B Additional Experiments

B.1 MuJoCo Control Suite

We evaluate our imitation learning approach, **RIZE**, on four benchmark Gym (Brockman et al., 2016) MuJoCo (Todorov et al., 2012) environments: *HalfCheetah-v2*, *Walker2d-v2*, *Ant-v2*, and *Humanoid-v2*. Expert trajectories were obtained from IQ-Learn (Garg et al., 2021), generated by a Soft Actor-Critic agent (Haarnoja et al., 2018), with each trajectory comprising 1,000 state-action transitions. To facilitate performance comparisons, episode returns are normalized relative to expert performance: *HalfCheetah* (5100), *Walker2d* (5200), *Ant* (4700), and *Humanoid* (5300).

B.2 Implementation Details

Our architecture integrates components from **Distributional SAC (DSAC)**[1] (Ma et al., 2020) and **IQ-Learn**[2] (Garg et al., 2021), with hyperparameters tuned through systematic search and ablation studies. Key configurations for experiments involving three and ten demonstrations are summarized in Table 1.

Distributional SAC Components: The critic network is implemented as a three-layer multilayer perceptron (MLP) with 256 units per layer, trained using a learning rate of 3×10^{-4}. The policy network is a four-layer MLP, also with 256 units per layer. Environment-specific learning rates are applied: 1×10^{-5} for the *Humanoid* environment and 5×10^{-5} for all others. To enhance training stability, we employ a target policy—a delayed version of the online policy—and sample next-state actions from this module. For value distribution training $Z^{\pi}_{\phi,\tau}$, we adopt the Implicit Quantile Networks (IQN) (Dabney et al., 2018a) approach by sampling quantile fractions τ uniformly from $\mathcal{U}(0, 1)$. Additionally, dual critic networks with delayed updates are used, which empirically improve training stability.

IQ-Learn Adaptations: Key adaptations from IQ-Learn include adjustments to the regularizer coefficient c and entropy coefficient α. Specifically, for the regularizer coefficient c, we find that $c = 0.5$ yields robust performance on the *Humanoid* task, while $c = 0.1$ works better for other tasks. For the entropy coefficient α, smaller values lead to more stable training. Unlike RL, where exploration is crucial, imitation learning relies less on entropy due to the availability of expert data. Across all tasks, we set initial target reward parameters as $\lambda^{\pi_E} = 10$ and $\lambda^{\pi} = 5$. Furthermore, we observe that lower learning rates for target rewards improve overall learning performance.

Previous implicit reward methods such as IQLearn, ValueDICE, and LSIQ[3] have employed distinct modifications to the loss function. In our setup, two main loss variants are defined:

- *value loss:*

$$\mathcal{L}(\pi, Q) = \mathbb{E}_{\rho_E}[Q(s, a) - \gamma V(s')] - \mathbb{E}_{\rho}[V(s) - \gamma V(s')] - c\,\Gamma(R_Q, \lambda)$$

- *v0 loss:*

$$\mathcal{L}(\pi, Q) = \mathbb{E}_{\rho_E}[Q(s, a) - \gamma V(s')] - (1 - \gamma)\mathbb{E}_{p_0}[V(s_0)] - c\,\Gamma(R_Q, \lambda)$$

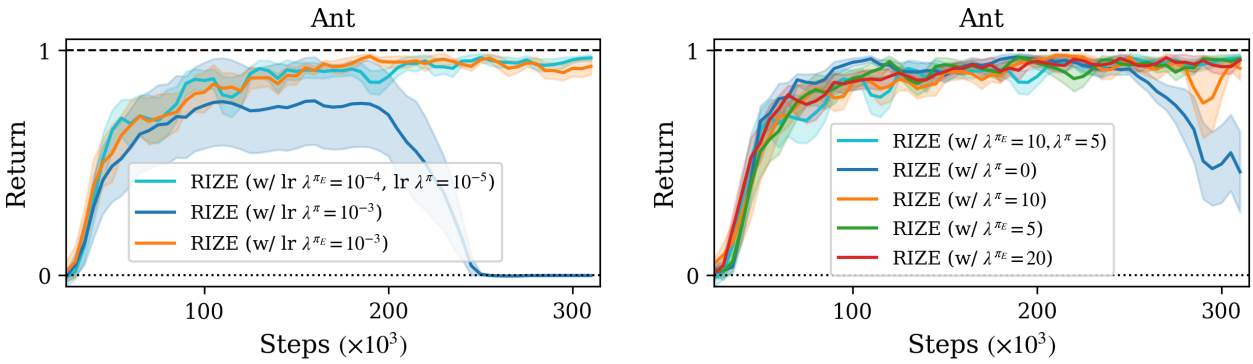
Here, ρ is a mixture distribution, p_0 denotes the initial distribution, $R_Q(s, a)$ is the implicit reward defined as $R_Q(s, a) = Q(s, a) - \gamma V(s')$, the state-value function is given by $V(s') = Q(s', a') - \alpha \log \pi(a'|s')$, and lastly, our convex regularizer is expressed as $\Gamma(R_Q, \lambda) = \mathbb{E}_{\rho_E}[(R_Q - \lambda^{\pi_E})^2] + \mathbb{E}_{\rho_\pi}[(R_Q - \lambda^{\pi})^2]$.

The choice between *v0* or *value* loss variants depends on environment complexity: we find that for a complex task like *Humanoid-v2*, the *v0* variant demonstrates greater robustness. And, for *HalfCheetah-v2*, *Walker2d-v2*, and *Ant-v2*, the *value* variant performs better.

[1] `https://github.com/xtma/dsac`
[2] `https://github.com/Div99/IQ-Learn`
[3] `https://github.com/robfiras/ls-iq/tree/main`

Table 1: Hyperparameters for 3 and 10 Demonstrations (merged where identical)

Environment	α (3 / 10)	c	lr π	lr Z	λ^{π_E}	lr λ^{π_E}	λ^{π}	lr λ^{π} (3 / 10)
Ant-v2	0.05 / 0.10	0.1	5×10^{-5}	3×10^{-4}	10	1×10^{-4}	5	1×10^{-5} / 1×10^{-4}
HalfCheetah-v2	0.05 / 0.10	0.1	5×10^{-5}	3×10^{-4}	10	1×10^{-4}	5	1×10^{-5} / 1×10^{-4}
Walker2d-v2	0.05 / 0.10	0.1	5×10^{-5}	3×10^{-4}	10	1×10^{-4}	5	1×10^{-5} / 1×10^{-4}
Humanoid-v2	0.05 / 0.10	0.5	1×10^{-5}	3×10^{-4}	10	1×10^{-4}	5	5×10^{-5} / 1×10^{-5}



(a) (b)

Figure 7: Fine-tuning analysis of target parameters. (a) Learning rates: Turquoise represents our method's primary result with learning rates of 1e−4 for λ^{π_E} and 1e−5 for λ^{π}. Orange and blue lines indicate higher learning rates (e.g., 1e−3) for λ^{π_E} and λ^{π}, respectively. (b) Starting values: Turquoise shows the main result with initial values of 10 for λ^{π_E} and 5 for λ^{π}, while other lines explore different starting values. Three trajectories are used throughout the analysis.

B.3 Hyperparameter Tuning

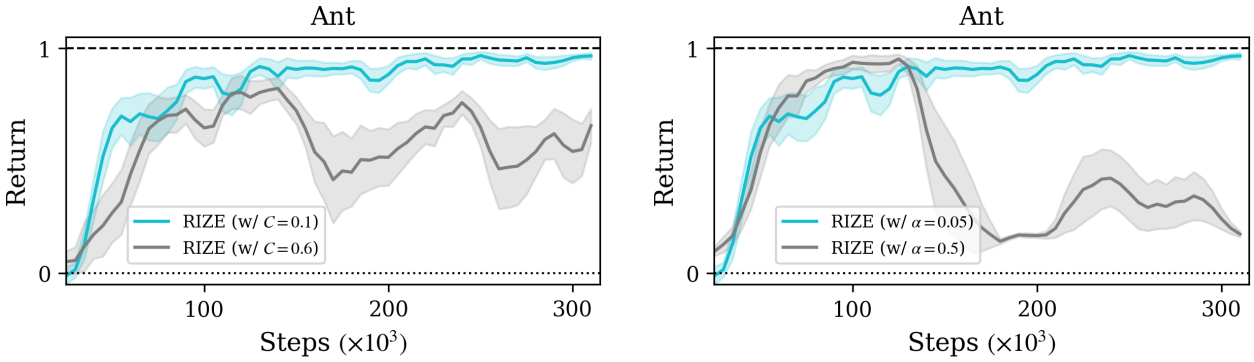
We present our analysis and comparison of important hyperparameters utilized in our algorithm. As before, all experiments use five seeds, and we show mean and half a standard deviation of all the seeds. As a challenging task among MuJoCo environments, we only experiment with Ant-v2.

Target Reward λ

Selecting appropriate initial values and learning rates for the automatic fine-tuning of λ^{π_E} and λ^{π} is critical in our approach. First, we observe that a suitable learning rate is essential for the stable training of our imitation learning agent, as illustrated in Figure 7a. Our findings indicate that λ^{π} must be optimized very slowly; using larger learning rates can destabilize training and hinder progress. In contrast, λ^{π_E} demonstrates greater resilience when optimized with higher learning rates. Additionally, λ^{π_E} remains robust even with varying initial values. However, as shown in Figure 7b, failing to select an appropriate initial value for λ^{π} can negatively impact learning. Overall, Figures 7a and 7b highlight the need for careful selection of both the learning rate and initial value when optimizing λ^{π}, while λ^{π_E} exhibits considerable robustness in this regard.

Regularization Coefficient c

Our experiments on the regularizer coefficient c reveal that smaller values of c encourage expert-like performance, while larger values overly constrain rewards and targets, limiting learning. This finding highlights the critical role of selecting an appropriate c, as it directly impacts the balance between learning from expert data and regularization: higher values prioritize regularization at the cost of learning, whereas smaller values favor learning but reduce regularization (see Figure 8a).



(a) (b)

Figure 8: (a) Effect of the regularizer coefficient c. Turquoise shows the primary result of our method with $c = 0.1$, while gray represents a larger value ($c = 0.6$). (b) Effect of the temperature parameter α. Turquoise shows the result with $\alpha = 0.05$, and gray corresponds to a larger value ($\alpha = 0.5$). Three trajectories are used throughout the analysis.

Entropy Coefficient α

We observe that the entropy coefficient is a crucial hyperparameter in inverse reinforcement learning (IRL) problems. As shown in Figure 8b, IRL methods typically require small values for α, a point previously noted (Garg et al., 2021). With expert demonstrations available, an imitation learning (IL) policy does not need to explore for optimal actions, as these are provided by the demonstrations. Consequently, higher values of α can lead to training instability, ultimately resulting in policy collapse.