INTERPRETING DISTRIBUTIONAL REINFORCEMENT LEARNING: A REGULARIZATION PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributional reinforcement learning (RL) is a class of state-of-the-art algorithms that estimate the entire distribution of the total return rather than its expected value alone. The theoretical advantages of distributional RL over expectation-based RL remain elusive, despite the remarkable performance of distributional RL. Our work attributes the superiority of distributional RL to its regularization effect stemming from the value distribution information regardless of only its expectation. We decompose the value distribution into its expectation and the remaining distribution part using a variant of the gross error model in robust statistics. Hence, distributional RL has an additional benefit over expectation-based RL thanks to the impact of a *risk-sensitive entropy regularization* within the Neural Fitted Z-Iteration framework. Meanwhile, we investigate the role of the resulting regularization in actor-critic algorithms by bridging the risk-sensitive entropy regularization of distributional RL and the vanilla entropy in maximum entropy RL. It reveals that distributional RL induces an augmented reward function, which promotes a risk-sensitive exploration against the intrinsic uncertainty of the environment. Finally, extensive experiments verify the importance of the regularization effect in distributional RL, as well as the mutual impacts of different entropy regularizations. Our study paves the way towards a better understanding of distributional RL, especially when looked at through a regularization lens.

1 INTRODUCTION

The intrinsic characteristics of classical reinforcement learning (RL) algorithms, such as temporal-difference (TD) learning (Sutton & Barto, 2018) and Q-learning (Watkins & Dayan, 1992), are based on the expectation of discounted cumulative rewards that an agent observes while interacting with the environment. In stark contrast to the classical expectation-based RL, a new branch of algorithms called distributional RL estimates the full distribution of total returns and has demonstrated the state-of-the-art performance in a wide range of environments (Bellemare et al., 2017a; Dabney et al., 2018b;a; Yang et al., 2019; Zhou et al., 2020; Nguyen et al., 2020; Sun et al., 2022). Meanwhile, distributional RL also inherits other benefits in risk-sensitive control (Dabney et al., 2018a), policy exploration settings (Mavrin et al., 2019; Rowland et al., 2019) and robustness (Sun et al., 2021).

Despite the existence of numerous algorithmic variants of distributional RL with remarkable empirical success, we still have a poor understanding of what the effectiveness of distributional RL is stemming from and theoretical studies on advantages of distributional RL over expectation-based RL are still less established. An existing work (Lyle et al., 2021) investigated the impact of distributional RL from the perspective of representation dynamics. Distributional RL problems was also mapped to a Wasserstein gradient flow problem (Martin et al., 2020), treating the distributional Bellman residual as a potential energy functional. Offline distributional RL (Ma et al., 2021) has also been proposed to investigate the efficacy of distributional RL in both risk-neutral and risk-averse domains. Although the explanation from these works is not sufficient yet, the trend is encouraging for recent works towards closing the gap between theory and practice in distributional RL.

In this paper, we illuminate the superiority of distributional RL over expectation-based RL through the lens of regularization to explain its empirical outperformance in most practical environments. Specifically, we simplify distributional RL into a *Neural Fitted Z-Iteration* framework, within which we establish an equivalence of objective functions between distributional RL and a risk-sensitive entropy regularized maximum likelihood estimation (MLE) from the perspective of statistics. This result is based on two analysis components, i.e., action-value distribution decomposition by leverage of a variant of gross error model in robust statistics, as well as Kullback-Leibler (KL) divergence to measure the distribution distance between the current and target value distribution in each Bellman update. Then we establish a connection between the impact of risk-sensitive entropy regularization of distributional RL and vanilla entropy in maximum entropy RL, yielding a *Distribution-Entropy-Regularized Actor Critic* algorithm. Empirical results demonstrate the crucial role of risk-sensitive entropy regularization effect from distributional RL in the superiority over expectation-based RL on both Atari games and MuJoCo environments, and reveal their mutual impacts of both risk-sensitive entropy in distributional RL and vanilla entropy in maximum entropy RL.

2 PRELIMINARY KNOWLEDGE

In classical RL, an agent interacts with an environment via a Markov decision process (MDP), a 5-tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively. P is the environment transition dynamics, R is the reward function and $\gamma \in (0,1)$ is the discount factor.

Action-value function vs Action-value distribution. Given a policy π, the discounted sum of future rewards is a random variable $Z^{\pi}(s,a) = \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)$, where $s_0 = s$, $a_0 = a$, $s_{t+1} \sim P(\cdot|s_t, a_t)$, and $a_t \sim \pi(\cdot|s_t)$. In the control setting, expectation-based RL focuses on the action-value function $Q^{\pi}(s,a)$, the expectation of $Z^{\pi}(s,a)$, i.e., $Q^{\pi}(s,a) = \mathbb{E}\left[Z^{\pi}(s,a)\right]$. Distributional RL, on the other hand, focuses on the action-value distribution, the full distribution of $Z^{\pi}(s,a)$. Leveraging knowledge on the entire distribution can better capture the intrinsic uncertainty of environment (Dabney et al., 2018a; Mavrin et al., 2019).

Bellman operators vs distributional Bellman operators. For the policy evaluation in expectation-based RL, the value function is updated via the Bellman operator $\mathcal{T}^{\pi} Q(s,a) = \mathbb{E}[R(s,a)] + \gamma \mathbb{E}_{s' \sim p, a' \sim \pi}\left[Q\left(s', a'\right)\right]$. In distributional RL, the action-value distribution of $Z^{\pi}(s,a)$ is updated via the distributional Bellman operator $\mathfrak{T}^{\pi}$

$$\mathfrak{T}^{\pi} Z(s,a) = R(s,a) + \gamma Z\left(s', a'\right), \tag{1}$$

where $s' \sim P(\cdot|s,a)$ and $a' \sim \pi\left(\cdot|s'\right)$. The equality in Eq. 1 implies that random variables of both sides are equal in distribution. This random-variable definition of distributional Bellman operator is appealing and easily understood due to its concise form, although its value-distribution definition is more mathematically rigorous (Rowland et al., 2018; Bellemare et al., 2022). More importantly, both the Bellman operator $\mathcal{T}^{\pi}$ and Bellman optimality operator $\mathcal{T}^{\text{opt}}$, defined as $\mathcal{T}^{\text{opt}} Q(s,a) = \mathbb{E}[R(s,a)] + \gamma \max_{a'} \mathbb{E}_{s' \sim p}\left[Q\left(s', a'\right)\right]$, in expectation-based RL are contractive in the stationary policy case. In contrast, the distributional Bellman operator $\mathfrak{T}^{\pi}$ is contractive under certain distribution divergences, but the distributional Bellman optimality operator can only converge to a set of optimal non-stationary value distributions in a weak sense (Elie & Arthur, 2020).

3 REGULARIZATION EFFECT OF DISTRIBUTIONAL RL

3.1 DISTRIBUTIONAL RL: NEURAL FITTED Z-ITERATION (NEURAL FZI)

Expectation-based RL: Neural Fitted Q-Iteration (Neural FQI). Neural FQI (Fan et al., 2020; Riedmiller, 2005) offers a statistical explanation of DQN (Mnih et al., 2015), capturing its key features, including experience replay and the target network Q_{θ^*}. In *Neural FQI*, we update parameterized $Q_{\theta}(s,a)$ in each iteration k in a regression problem:

$$Q_{\theta}^{k+1} = \underset{Q_{\theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left[y_i - Q_{\theta}^{k}\left(s_i, a_i\right)\right]^2, \tag{2}$$

where the target $y_i = r(s_i, a_i) + \gamma \max_{a \in \mathcal{A}} Q_{\theta^*}^{k}\left(s_i', a\right)$ is fixed within every T_{target} steps to update target network Q_{θ^*} by letting $\theta^* = \theta$. The experience buffer induces independent samples $\left\{\left(s_i, a_i, r_i, s_i'\right)\right\}_{i \in [n]}$. In an ideal case when we neglect the non-convexity and TD approximation errors, we have $Q_{\theta}^{k+1} = \mathcal{T}^{\text{opt}} Q_{\theta}^{k}$, which is exactly the updating rule under Bellman optimality operator (Fan et al., 2020). In the viewpoint of statistics, the optimization problem in Eq. 2 can be viewed as Least Square Estimation (LSE) in a neural network parametric regression problem regarding Q_{θ}.

Distributional RL: Neural Fitted Z-Iteration (Neural FZI). *We interpret distributional RL as a Neural Fitted Z-Iteration owing to the fact that this iteration is by far closest to the practical algorithms and more interpretable.* Analogous to Neural FQI, we can simplify value-based distributional RL algorithms parameterized by Z_θ into a *Neural Fitted Z-Iteration (Neural FZI)* as

$$Z_\theta^{k+1} = \underset{Z_\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} d_p(Y_i, Z_\theta^k(s_i, a_i)), \tag{3}$$

where the target $Y_i = R(s_i, a_i) + \gamma Z_{\theta^*}^k(s_i', \pi_Z(s_i'))$ with the policy π_Z following the greedy rule $\pi_Z(s_i') = \operatorname{argmax}_{a'} \mathbb{E}\left[Z_{\theta^*}^k(s_i', a')\right]$ is fixed within every T_{target} steps to update target network Z_{θ^*}. d_p is a divergence between two distributions. Notably, *choices of representation for Z_θ and the metric d_p are pivotal for the empirical success of distributional RL algorithms.* For instance, QR-DQN (Dabney et al., 2018b) leverages quantiles to represent the distribution of Z_θ and approximates Wasserstein distance via quantile regression. C51 (Bellemare et al., 2017a) employs a categorical parameterization on Z_θ and select Kullback–Leibler (KL) divergence as d_p, whose convergence has been shown under the Cramér distance (Bellemare et al., 2017b; Rowland et al., 2018). Moment Matching DQN (Nguyen et al., 2020) learns deterministic samples to express the distribution of Z_θ and optimize based on Maximum Mean Discrepancy (MMD), while the recent Sinkhorn distributional RL (Sun et al., 2022) is based on the "intermediate" Sinkhorn divergence as d_p.

3.2 DISTRIBUTIONAL RL: ENTROPY-REGULARIZED MLE IN NEURAL FZI

In order to provide a statistical interpretation about the superiority of distributional RL as opposed to expectation-based RL, we rewrite the objective function in Neural FZI framework of distributional RL as an entropy regularized Maximum Likelihood Estimation (MLE) from the viewpoint of statistics *under the action-value distribution decomposition and KL divergence as d_p.*

Analysis Component 1: Action-Value Distribution Decomposition. To separate the impact of additional distribution information from the expectation of Z^π, we leverage a variant of *gross error model* from robust statistics (Huber, 2004), which was also similarly used to analyze Label Smoothing (Müller et al., 2019) and Knowledge Distillation (Hinton et al., 2015). Specifically, we denote the one-dimensional full cumulative distribution function (cdf) of $Z^\pi(s, a)$ as $F^{s,a}$, and assume that this action-value distribution $F^{s,a}$ satisfies the following expectation decomposition:

$$F^{s,a}(x) = (1 - \epsilon)\mathbb{1}_{\{x \geq \mathbb{E}[Z^\pi(s,a)]\}}(x) + \epsilon F_\mu^{s,a}(x), \tag{4}$$

where $F_\mu^{s,a}$ can be determined by $F^{s,a}$ and the specified ϵ, aiming at characterizing the impact of action-value distribution *regardless of* its expectation $\mathbb{E}[Z^\pi(s, a)]$. ϵ controls the proportion of $F_\mu^{s,a}(x)$ and the indicator function $\mathbb{1}_{\{x \geq \mathbb{E}[Z^\pi(s,a)]\}} = 1$ if $x \geq \mathbb{E}[Z^\pi(s, a)]$, otherwise 0. Although we can let $F_\mu^{s,a}$ be an arbitrarily continuous and differential distribution, the function class of $F^{s,a}$ is slightly restricted even discontinuous in order to satisfy the expectation decomposition in Eq. 4. Nevertheless, in Proposition 1 we show that for an arbitrarily continuous cumulative distribution function F, the distance between F and $F^{s,a}$ in the supreme norm can be upper bounded under mild assumptions, indicating the rationale of value distribution restriction on $F^{s,a}$ proposed in Eq. 4.

Proposition 1. *Given an arbitrarily continuous random variable $Z^\pi(s, a)$ with the distribution function F and expectation $\mathbb{E}[Z^\pi(s, a)]$, it holds that:*

$$\inf_{F_\mu^{s,a}} \|F - F^{s,a}\|_\infty \leq (1 - \epsilon) \max\{1 - F(\mathbb{E}[Z^\pi(s, a)]), F(\mathbb{E}[Z^\pi(s, a)])\}. \tag{5}$$

If Z^π is bounded in $[-c, c]$ and has variance σ^2, we have: $\inf_{F_\mu^{s,a}} \|F - F^{s,a}\|_\infty \leq (1-\epsilon)(1-\frac{\sigma^2}{2c^2})$.

The proof of Proposition 1 is provided in Appendix A. In particular, $F^{s,a}$ will converge to F uniformly over x when $\epsilon \to 1$. *We highlight that $F_\mu^{s,a}$ would be central to the regularization analysis in distributional RL.* Under the distribution decomposition in Eq. 4, we immediately attain their density function relationship as $p^{s,a}(x) = (1 - \epsilon)\delta_{\{x=\mathbb{E}[Z^\pi(s,a)]\}}(x) + \epsilon \mu^{s,a}(x)$, where $\delta_{\{x=\mathbb{E}[Z^\pi(s,a)]\}}$ is a Dirac function centered at $\mathbb{E}[Z^\pi(s, a)]$. $\mu^{s,a}(x)$ is the probability density function (pdf) of $F_\mu^{s,a}$ related to $Z^\pi(s, a)$ that depends upon $\mathbb{E}[Z^\pi(s, a)]$. Next, we use $p^{s,a}(x)$ to express the true target probability density function behind $\{Y_i\}_{i\in[n]}$, and $q_\theta^{s,a}(x)$ to denote the approximated one of $Z_\theta^k(s, a)$ in Neural FZI in Eq. 3.

Analysis Component 2: Kullback–Leibler (KL) Divergence as d_p in Neural FZI. We find it difficult to directly focus on the more commonly used Wasserstein distance to conduct a theoretical analysis, and instead we shift our attention to KL divergence. Our motivations are multiple:

- As a widely-used divergence to distribution distance, the KL divergence is also successfully applied in categorical distributional RL, e.g., C51 (Bellemare et al., 2017a), that can be viewed as the first successful distributional RL algorithm.
- The choice of KL divergence will establish a theoretical connection between distributional RL and maximum entropy RL, e.g., Soft Q-Learning (Haarnoja et al., 2017) and Soft Actor Critic (SAC) (Haarnoja et al., 2018).
- In Proposition 2 (proof is given in Appendix B), we summarize that KL divergence enjoys desirable properties in distributional RL context, including a non-expansion distributional Bellman operator, a close link with Wasserstein distance and the expectation contraction property. As such, KL divergence can be reasonably used for the theoretical analysis.

Proposition 2. *Given two probability measures μ and ν, we define the supreme D_{KL} as a functional $\mathcal{P}(\mathcal{X})^{\mathcal{S}\times\mathcal{A}} \times \mathcal{P}(\mathcal{X})^{\mathcal{S}\times\mathcal{A}} \rightarrow \mathbb{R}$, i.e., $D_{KL}^{\infty}(\mu, \nu) = \sup_{(x,a)\in\mathcal{S}\times\mathcal{A}} D_{KL}(\mu(x,a), \nu(x,a))$. we have: (1) $\mathfrak{T}^{\pi}$ is a non-expansive distributional Bellman operator under D_{KL}^{∞}, i.e., $D_{KL}^{\infty}(\mathfrak{T}^{\pi} Z_1, \mathfrak{T}^{\pi} Z_2) \leq D_{KL}^{\infty}(Z_1, Z_2)$, (2) $D_{KL}^{\infty}(Z_n, Z) \rightarrow 0$ implies the Wasserstein distance $W_p(Z_n, Z) \rightarrow 0$, (3) the expectation of Z^{π} is still γ-contractive under D_{KL}^{∞}, i.e., $\left\|\mathbb{E}\mathfrak{T}^{\pi} Z_1 - \mathbb{E}\mathfrak{T}^{\pi} Z_2\right\|_{\infty} \leq \gamma \left\|\mathbb{E}Z_1 - \mathbb{E}Z_2\right\|_{\infty}$.*

Putting Them Together: Entropy-regularized MLE for Distributional RL. We incorporate both value distribution decomposition and KL divergence as d_p in Neural FZI (Eq. 3) of distributional RL. Let $\mathcal{H}(P, Q)$ be the cross entropy between two probability measures P and Q, i.e., $\mathcal{H}(P, Q) = -\int_{x\in\mathcal{X}} P(x) \log Q(x) \,\mathrm{d}x$. We can derive the following entropy-regularized MLE form for distributional RL in Proposition 3. Please refer to the proof in Appendix C.

Proposition 3. *Denote α as a positive constant, and based on the value distribution decomposition in Eq. 4 and D_{KL} as d_p, Neural FZI in Eq. 3 can be explicitly expressed as*

$$Z_{\theta}^{k+1} = \underset{q_{\theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left[-\log q_{\theta}^{s_i,a_i}\left(\mathbb{E}\left[Z^{\pi}(s_i', \pi_Z(s_i'))\right]\right) + \alpha\mathcal{H}(\mu^{s_i',\pi_Z(s_i')}, q_{\theta}^{s_i,a_i}) \right], \tag{6}$$

where $\alpha = \varepsilon/(1-\varepsilon) > 0$ and we use q_{θ} to denote the probability density function of Z_{θ}^{k} in the k-th iteration for conciseness. For the uniformity of notation, we still use s, a in the following analysis instead of s_i, a_i. Concretely, we find these two intriguing terms in Eq. 6 together serve as an entropy-regularized MLE and we call their impacts on the RL optimization as **expectation effect** and **distributional regularization effect**, respectively.

- For the **expectation effect** of the first term, using the language of statistics, minimizing it is equivalent to a variant of MLE over $q_{\theta}^{s,a}$ for the *current state-action pair* (s, a) on the **expectation** $\mathbb{E}\left[Z^{\pi}(s_i', \pi_Z(s_i'))\right]$ of the target action-value distribution for the *next state-action pair* $(s', \pi_Z(s_i'))$, leading to a similar optimization impact of expectation-based RL. This is in contrast to the classical MLE directly on *observed samples*, i.e., $Y_i \sim F^{s_i',\pi_Z(s_i')}$.
- For the **distributional regularization effect** of the second term, it pushes $q_{\theta}^{s,a}$ for the current state-action pair to approximate $\mu^{s_i',\pi_Z(s_i')}$ for the next state-action pair, which "deducts" the expectation effect from the whole action-value distribution by leverage of the value distribution decomposition in Eq. 4. Therefore, this regularization term serves as a crucial factor to interpret the advantage of distributional RL over expectation-based RL.

Risk-Sensitive Entropy Regularization. We attribute the superiority of distributional RL to significantly reduce intrinsic uncertainty of the environment (Mavrin et al., 2019) into the regularization term in Eq. 6. According to the literature of risks in RL (Dabney et al., 2018a), where "risk" refers to the uncertainty over possible outcomes and "risk-sensitive policies" are those which depend upon more than the mean of the outcomes, we hereby call the novel cross entropy regularization for the second term in Eq. 6 as *risk-sensitive entropy regularization*. This risk-sensitive entropy regularization derived within distributional RL expands the class of policies using information provided by the distribution over returns (i.e. to the class of risk-sensitive policies). It should also be noted that our risk-sensitive entropy regularization is indeed “risk-neural” in the sense of convexity or concaveness of utility functions, where our policy is still applying a linear utility function U, defined as $\pi(\cdot|s) = \arg\max_a \mathbb{E}_{Z(s,a)}[U(z)]$. Correspondingly, We can additionally vary different distortion risk measures to explicitly lead the policy to being risk-averse or risk-seeking (Dabney et al., 2018a).

Remark on the Attainability of $\mu^{s',\pi_Z(s')}$. In practical distributional RL algorithms, we typically use the bootstrap, e.g., TD learning, to attain the target distribution estimate $F^{s',\pi_Z(s')}$ and thus immediately obtain $\mu^{s',\pi_Z(s')}$ based on Eq. 4 as long as $\mathbb{E}\left[Z(s,a)\right]$ exists. The leverage of $\mu^{s',\pi_Z(s')}$ and the regularization effect revealed in Eq. 6 of distributional RL de facto establishes a bridge with maximum entropy RL (Williams & Peng, 1991), on which we have a deeper analysis in Section 3.3.

3.3 CONNECTION WITH MAXIMUM ENTROPY RL

Vanilla Entropy Regularization in Maximum Entropy RL. Maximum entropy RL (Williams & Peng, 1991), including Soft Q-Learning (Haarnoja et al., 2017), explicitly optimizes for policies that aim to reach states where they will have high entropy in the future:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}\left[r\left(s_t,a_t\right) + \beta\mathcal{H}(\pi(\cdot|s_t))\right], \tag{7}$$

where $\mathcal{H}\left(\pi_\theta\left(\cdot|s_t\right)\right) = -\sum_a \pi_\theta\left(a|s_t\right)\log\pi_\theta\left(a|s_t\right)$ and ρ_π is the generated distribution following π. The temperature parameter β determines the relative importance of the entropy term against the cumulative rewards, and thus controls the action diversity of the optimal policy learned via Eq. 7. This maximum entropy regularization has various conceptual and practical advantages. Firstly, the learned policy is encouraged to visit states *with high entropy in the future*, thus promoting the exploration over diverse states (Han & Sung, 2021). Secondly, it considerably improves the learning speed (Mei et al., 2020) and therefore is widely used in state-of-the-art algorithms, e.g., Soft Actor-Critic (SAC) (Haarnoja et al., 2018). Similar empirical benefits of both distributional RL and maximum entropy RL also encourage us to probe their underlying connection.

Risk-Sensitive Entropy Regularization in Distributional RL. To make a direct comparison with maximum entropy RL, we need to specifically analyze the impact of the regularization term in Eq. 6, and thus we incorporate the risk-sensitive entropy regularization of distributional RL into the policy gradient framework akin to maximum entropy RL. Concretely, we conduct our analysis by showing the convergence of *Distribution-Entropy-Regularized Policy Iteration*, which is counterpart for Soft Policy Iteration (Haarnoja et al., 2018), i.e., the underpinning of SAC algorithm. In principle, Distribution-Entropy-Regularized Policy Iteration replaces the vanilla entropy regularization in Soft Policy Iteration with our risk-sensitive entropy regularization in Eq. 6 from distributional RL. In the policy evaluation step of distribution-entropy-regularized policy iteration, a new soft Q-value, i.e., the expectation of $Z^\pi(s,a)$, can be computed iteratively by applying a modified Bellman operator $\mathcal{T}_d^\pi$, which we call *Distribution-Entropy-Regularized Bellman Operator* defined as

$$\mathcal{T}_d^\pi Q\left(s_t,a_t\right) \triangleq r\left(s_t,a_t\right) + \gamma\mathbb{E}_{s_{t+1}\sim\rho^\pi}\left[V\left(s_{t+1}|s_t,a_t\right)\right], \tag{8}$$

where a new soft value function $V\left(s_{t+1}|s_t,a_t\right)$ conditioned on s_t, a_t is defined by

$$V\left(s_{t+1}|s_t,a_t\right) = \mathbb{E}_{a_{t+1}\sim\pi}\left[Q\left(s_{t+1},a_{t+1}\right)\right] + f(\mathcal{H}\left(\mu^{s_t,a_t}, q_\theta^{s_t,a_t}\right)), \tag{9}$$

and f is a continuous *increasing* function over the cross entropy $\mathcal{H}$. Note that in this specific tabular setting regarding s_t and a_t, we particularly use $q_\theta^{s_t,a_t}(x)$ to approximate the true density function of $Z(s_t,a_t)$, and μ^{s_t,a_t} to represent the target value distribution *regardless of* its expectation, which can normally be obtained via bootstrap estimate $\mu^{s_{t+1},\pi_Z(s_{t+1})}$ similar in Eq. 6. The f transformation over the cross entropy $\mathcal{H}$ between μ^{s_t,a_t} and $q_\theta^{s_t,a_t}(x)$ serves as our *risk-sensitive entropy regularization*. As opposed to the vanilla entropy regularization in maximum entropy RL that encourages the policy to explore, our risk-sensitive entropy regularization in distributional RL plays a role of *the reward correction* or *augmented reward*, and therefore augments the action-value function $Q(s_t,a_t)$ in the value-based RL and the objective function in policy gradient RL by additionally incorporating the value distribution knowledge. As we have discussed Neural FZI above in Section 3.2, which is established on the value-based RL, we now shift our attention to the properties of our risk-sensitive entropy regularization in the framework of policy gradient. In Lemma 1, we firstly show that our Distribution-Entropy-Regularized Bellman operator $\mathcal{T}_d^\pi$ still inherits the convergence property in the policy evaluation phase.

Lemma 1. *(Distribution-Entropy-Regularized Policy Evaluation) Consider the distribution-entropy-regularized Bellman operator* $\mathcal{T}_d^\pi$ *in Eq. 8 and the behavior of expectation of* $Z^\pi(s,a)$*, i.e.,* $Q(s,a)$*. Assume* $\mathcal{H}^\pi(\mu^{s_t,a_t}, q_\theta^{s_t,a_t}) \leq M$ *for all* $(s_t,a_t) \in \mathcal{S} \times \mathcal{A}$*, where* M *is a constant. Define* $Q^{k+1} = \mathcal{T}_{sd}^\pi Q^k$*, then* Q^{k+1} *will converge to a corrected Q-value of* π *as* $k \to \infty$ *with the new objective function defined as*

$$J'(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}\left[r(s_t,a_t) + \gamma f(\mathcal{H}(\mu^{s_t,a_t}, q_\theta^{s_t,a_t}))\right]. \tag{10}$$

In Lemma 1, we reveal that the new objective function for distributional RL can be interpreted as an augmented reward function. Secondly, in the policy improvement for distributional RL, we keep the vanilla policy improvement updating rules according to

$$\pi_{\text{new}} = \arg\max_{\pi' \in \Pi} \mathbb{E}_{a_t\sim\pi'}\left[Q^{\pi_{\text{old}}}(s_t,a_t)\right]. \tag{11}$$

Next we can immediately derive a new policy iteration algorithm, called *Distribution-Entropy-Regularized Policy Iteration* that alternates between distribution-entropy-regularized policy evaluation in Eq. 8 and the policy improvement in Eq. 11. It will provably converge to the policy with the optimal risk-sensitive entropy among all policies in Π as shown in Theorem 1.

Theorem 1. *(Distribution-Entropy-Regularized Policy Iteration) Assume* $\mathcal{H}^\pi(\mu^{s_t,a_t}, q_\theta^{s_t,a_t}) \leq M$ *for all* $(s_t,a_t) \in \mathcal{S} \times \mathcal{A}$*, where* M *is a constant. Repeatedly applying distribution-entropy-regularized policy evaluation in Eq. 8 and the policy improvement in Eq. 11, the policy converges to an optimal policy* π^* *such that* $Q^{\pi^*}(s_t,a_t) \geq Q^\pi(s_t,a_t)$ *for all* $\pi \in \Pi$.

Please refer to Appendix D for the proof of Lemma 1 and Theorem 1. According to Theorem 1, it turns out that if we incorporate the risk-sensitive entropy regularization into the policy gradient framework in Eq. 10, we are able to design a variant of "soft policy iteration" that can guarantee the convergence to an optimal policy. As such, we provide a comprehensive comparison between vanilla entropy in maximum entropy RL and risk-sensitive entropy in distributional RL as follows.

Vanilla Entropy Regularization vs Risk-Sensitive Entropy Regularization. (1) Objective function. By comparing two objective function $J(\pi)$ in Eq. 7 for maximum entropy RL and $J'(\pi)$ in Eq. 10 for distributional RL, distributional RL tries to maximize the risk-sensitive entropy regularization *w.r.t.* π. This indicates that the learned policy in distributional RL is encouraged to *visit state and action pairs in the future whose action-value distributions have a higher degree of dispersion, e.g., variance, in spite of its expectation*, thus promoting the **risk-sensitive exploration** to reduce the intrinsic uncertainty of the environment. An intuitive illustration is provided in Figure 1. **(2) State-action dependent regularization.** The vanilla entropy $\mathcal{H}(\pi(\cdot|s_t))$ in maximum entropy RL is state-wise, while our risk-sensitive regularization $\mathcal{H}(\mu^{s_t,a_t}, q_\theta^{s_t,a_t})$ is state-action-wise, implying that it is a more fine-grained regularization to characterize the action-value distribution of $Z(s_t,a_t)$ in the future.

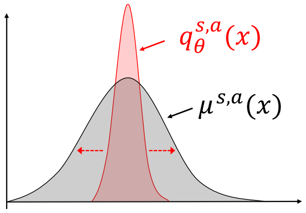

Figure 1: Impact of the risk-sensitive entropy regularization in distributional RL.

3.4 ALGORITHM: DISTRIBUTION-ENTROPY-REGULARIZED ACTOR-CRITIC (DERAC)

In practice, large continuous domains require us to derive a practical approximation to Distribution-Entropy-Regularized Policy Iteration (DERPI). We thus extend DERPI from the tabular setting to the function approximation case, yielding the Distribution-Entropy-Regularized Actor-Critic (DERAC) algorithm by using function approximators for both the value distribution $q_\theta(s_t,a_t)$ and the policy $\pi_\phi(a_t|s_t)$. The key characteristics of DERAC algorithm is that we use function approximator to represent the whole value distribution q_θ rather than only the value function, and conduct the optimization mainly based on the value function $Q_\theta(s_t,a_t) = \mathbb{E}\left[q_\theta(s_t,a_t)\right]$.

Optimize the parameterized value distribution q_θ. The new value function is originally trained to minimize the squared residual error of Eq. 8. Here for a desirable interpretation, we impose the zero expectation assumption over the residual, i.e., $\mathcal{T}^\pi Q_\theta(s,a) = Q_\theta(s,a) + \epsilon$ with $\mathbb{E}[\epsilon] = 0$. The resulting simplified objection function $\hat{J}_q(\theta)$ can be well interpreted as an interpolation between the expectation effect and distributional regularization effect:

$$\begin{aligned}\hat{J}_q(\theta) &= \mathbb{E}_{s,a}\left[\left(\mathcal{T}_d^\pi Q_\theta(s,a) - Q_\theta(s,a)\right)^2\right] \\ &\propto (1-\lambda)\mathbb{E}_{s,a}\left[(\mathcal{T}^\pi \mathbb{E}\left[q_\theta(s,a)\right]] - \mathbb{E}\left[q_\theta(s,a)\right])^2 + \lambda \mathbb{E}_{s,a}\left[\mathcal{H}(\mu^{s,a}, q_\theta^{s,a})\right],\end{aligned} \tag{12}$$

where we consider a particular increasing function $f(\mathcal{H}) = (\lambda\mathcal{H})^{\frac{1}{2}}/\gamma$ and $\lambda \in [0,1]$ is the hyper-parameter that controls the risk-sensitive regularization effect. Interestingly, when we leverage the whole target value distribution $F^{s,a}$ to approximate the true $\mu^{s,a}$, the objective function in Eq. 12 can be viewed as an exact interpolation of loss functions between expectation-based RL (the first term) and distributional RL equipped with KL divergence (the second term), e.g., C51. Note that for the target $\mathcal{T}^\pi \mathbb{E}[q_\theta(s,a)]$, we use the target value distribution neural network q_{θ^*} to stabilize the training, which is consistent with the Neural FZI framework analyzed in Section 3.1.

Optimize the policy π_ϕ. We optimize π_ϕ in Eq. 11 based on the $Q(s,a)$ and thus the new objective function $\hat{J}_\pi(\phi)$ can be expressed as $\hat{J}_\pi(\phi) = \mathbb{E}_{s,a\sim\pi_\phi}\left[\mathbb{E}\left[q_\theta(s,a)\right]\right]$. The complete DERAC algorithm is described in Algorithm 1 of Appendix F.

4 EXPERIMENTS

In the experiment, we firstly verify the regularization effect of distributional RL analyzed in Section 3.2 by decomposing the value distribution via Eq. 6 on both Atari games and MuJoCo environments. Next, we demonstrate the performance of DERAC algorithm on continual control environments. Finally, an empirical extension to Quantile-Regression Distributional RL, i.e., QR-DQN, is also provided to reveal mutual impacts of different entropy regularizations.

Environments. To demonstrate the value distribution decomposition, we mainly present results on three Atari games, including Breakout, Seaquest, Asterix, over 3 seeds and three continuous control MuJoCo environments in OpenAI Gym, including ant, swimmer and bipedalwalkerhardcore, over 5 seeds. For the extension to QR-DQN, we perform experiments on eight MuJoCo environments.

Baselines. To evaluate the risk-sensitive entropy regularization effect of distributional RL, we conduct an ablation study on C51 (Bellemare et al., 2017a) on Atari games and distributional SAC (DSAC) (Ma et al., 2020) on MuJoCo environments. The implementation of DERAC algorithm is based on distributional SAC (Haarnoja et al., 2018; Ma et al., 2020).

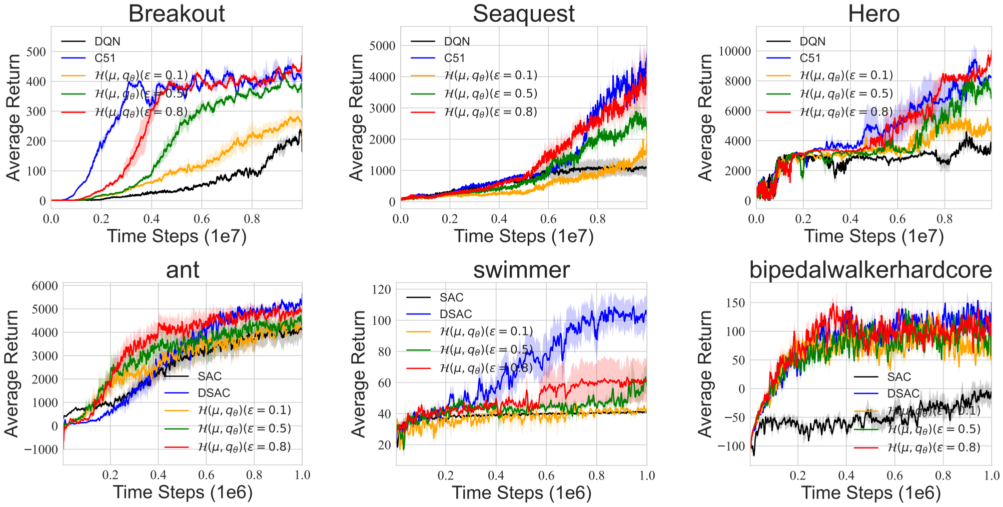

Figure 2: (**First Row**) Learning curves of C51 with value distribution decomposition $\mathcal{H}(\mu, q_\theta)$ under different ε on three Atari games over 3 seeds. (**Second Row**) Learning curves of C51 with value distribution decomposition $\mathcal{H}(\mu, q_\theta)$ under different ε on three MuJoCo environments over 5 seeds.

4.1 DISTRIBUTION REGULARIZATION EFFECT OF DISTRIBUTIONAL RL

We demonstrate the rationale of value distribution decomposition in Eq. 4 and the distribution regularization effect analyzed in Eq. 6 based on C51 algorithm equipped with KL divergence. Firstly, it is a fact that the value distribution decomposition is based on the equivalence between KL divergence and cross entropy owing to the leverage of target networks. Therefore, we firstly demonstrate that C51 algorithm can still achieve similar results *under the cross entropy loss* across four Atari games in Figure 5 of Appendix G. As C51 utilizes discrete histogram density function to approximate the true probability function, we extend the continuous decomposition $p^{s,a}(x) = (1-\epsilon)\delta_{\{x=\mathbb{E}[Z^\pi(s,a)]\}}(x) + \epsilon\mu^{s,a}(x)$ to a discrete form by decomposing the target value density function into an indicator function centered on the bin that contains the expectation and the remaining density function $\mu^{s,a}(x)$. We instead replace the true target probability $p^{s,a}(x)$ with $\mu^{s,a}(x)$ under different ε in the cross entropy loss, allowing to investigate the risk-sensitive regularization effect of distributional RL. Concretely, with a slight abuse of notation, we redefine ε as the proportion of probability of the bin that contains the expectation *with the mass to transport to other bins*. This is because an overly large true ε in Eq. 4 will result in a negative μ in some bins, which violates the definition of probability functions. On this account, we leverage the proportion probability ε rather than the true ε. Note that the new ε is still proportional to the true ε as a large proportion probability ε will transport less mass to other bins. This implies that the resulting $\mu^{s,a}(x)$ would be closer to the true probability $p^{s,a}(x)$, corresponding to a higher risk-sensitive regularization effect as analyzed in Eq. 6.

As shown in Figure 2, when ε gradually decreases from 0.8 to 0.1, the learning curves of C51 $\mathcal{H}(\mu, q_\theta)$ tend to degrade from vanilla C51 to DQN across both Atari and MuJoCo, although their sensitivity in terms of ε may depend on the environment, e.g., bipedalwalkerhardcore. This empirical observation corroborates the theoretical results we derive in Section 3.2, suggesting that risk-sensitive entropy regularization is pivotal to the success of distributional RL algorithms.

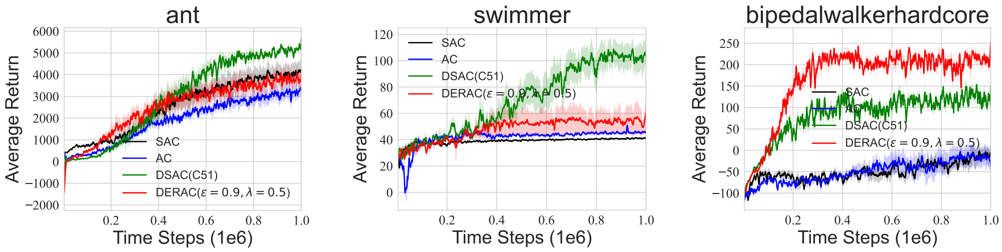

Figure 3: Learning curves of DERAC algorithms on three MuJoCo environments over 5 seeds.

4.2 CONVERGENCE OF DERAC ALGORITHM

We further demonstrate the convergence of *Distribution-Entropy-Regularized Actor-Critic (DERAC)*. Figure 3 showcases that DERAC is able to converge and achieve desirable performance on these three MuJoCo environments compared with AC (SAC without vanilla entropy) in the blue line. More importantly, Distribution-Entropy-Regularization (DER) in the red line could be remarkably beneficial for learning on the complex Bipedalwalkerhardcore, where a risk-sensitive exploration significantly improves the performance. It is worthwhile to know that our goal to introduce DERAC algorithm is not to pursue the empirical superiority of performance, but to corroborate the theoretical convergence of DERAC algorithm and DERPI in Theorem 1. Our empirical result in Figure 3 has provided strong evidence to verify our theoretical results. In addition, as we choose $\varepsilon = 0.9$ in DERAC algorithm, there exists a distribution information loss, resulting in the learning performance degradation. In practice, we can directly deploy distributional SAC to seek for a better performance. We also provide a sensitivity analysis of DERAC regarding λ in Figure 6 of Appendix G.

4.3 EXTENSION TO QUANTILE-REGRESSION DISTRIBUTIONAL RL

Finally, due to the fact that our aforementioned theoretical analysis is established on the distributional RL algorithms equipped with KL divergence, e.g., C51, in order to make a comprehen-

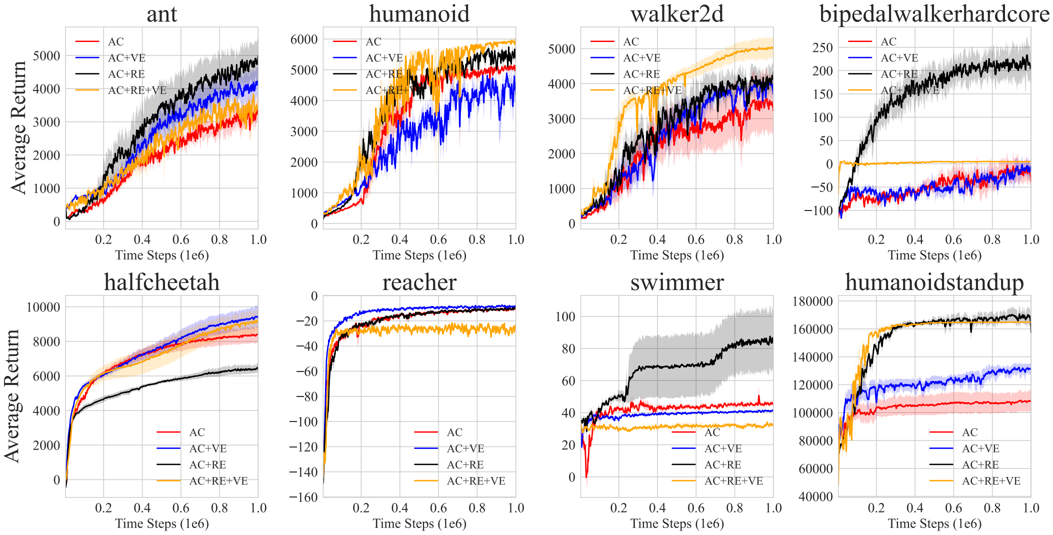

Figure 4: Learning curves of AC, AC+VE, AC+RE and AC+RE+VE over 5 seeds with smooth size 5 across eight MuJoCo environments where distributional RL part is based on IQN.

sive conclusion in broader distributional RL branches, we thus heuristically extend our results in quantile-regression-based distributional RL. Specifically, a careful ablation study is conducted to control the effects of vanilla entropy (VE), risk-sensitive entropy (RE) and their mutual impact. We denote SAC with and without vanilla entropy as *AC* and *AC+VE*, and distributional SAC with and without vanilla entropy as *AC+RE+VE* and *AC+RE*, where VE and RE are short for *Vanilla Entropy* and *Risk-sensitive Entropy*. For the implementation, we leverage the quantiles generation strategy in IQN (Dabney et al., 2018a) in distributional SAC (Ma et al., 2020). Hyper-parameters are listed in Appendix E. As suggested in Figure 4, although both vanilla entropy and risk-sensitive entropy effects may vary for different environments, we make the following conclusions:

(1) Vanilla entropy effect can enhance the performance as it is easily observed that AC+VE (blue line) outperforms AC (red lines) across most environments except on the humanoid and swimmer. The risk-sensitive entropy effect (RE) from distributional RL is also able to benefit the learning due to the fact that AC+RE (black lines) is more likely to bypass AC (red lines) especially on the complex BipealWalkerHardcore environment (hard for exploration).

(2) The use of both risk-sensitive entropy and vanilla entropy may interfere with each other, e.g., on BipealWalkerHardcore and Swimmer games, where *AC+RE+VE* (orange line) is significantly inferior to *AC+RE* (black line). This is because SAC encourages the policy to visit states with high entropy to pursue the diversity of states to optimize, while distributional RL promotes the risk-sensitive exploration to visit state and action pairs whose action-value distribution has more degree of dispersion. We hypothesize that these two different regularization effects are likely to lead to divergent optimization paths to optimize the policy for different exploration, e.g, gradient directions, thus interfering with each other eventually.

5 DISCUSSIONS AND CONCLUSION

Our regularization interpretation of distributional RL is mainly established on distributional RL with the KL divergence as d_p, while a direct analysis based on Wasserstein distance is also promising, albeit being theoretically tricky.

In this paper, we illuminate the superiority of distributional RL over expectation-based RL from the perspective of regularization. A risk-sensitive entropy regularization in the objective function is derived for distributional RL within Neural Fitted Z-Iteration to explain the benefit of distributional RL. Further, we also establish a connection between distributional RL with maximum entropy RL. Our research contributes to a deeper understanding of the advantage of distributional RL algorithms.

Ethics Statement. As our study is related to reveal the regularization effect of distributional RL algorithms, it is not involved with any ethics issue in our opinion.

Reproducibility Statement. Our results is based on the public implementation released in (Ma et al., 2020) with necessary implementation details given in Appendix E. We also provide the detailed proof from Appendix A to Appendix D.

REFERENCES

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *International Conference on Learning Representations, 2017*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. *International Conference on Machine Learning (ICML)*, 2017a.

Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017b.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2022. `http://www.distributional-rl.org`.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. *International Conference on Machine Learning (ICML)*, 2018a.

Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018b.

Odin Elie and Charpentier Arthur. *Dynamic Programming in Distributional Reinforcement Learning*. PhD thesis, Université du Québec à Montréal, 2020.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pp. 1352–1361. PMLR, 2017.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Seungyul Han and Youngchul Sung. A max-min entropy framework for reinforcement learning. *Advances in neural information processing systems (NeurIPS)*, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Peter J Huber. *Robust Statistics*, volume 523. John Wiley & Sons, 2004.

Clare Lyle, Mark Rowland, Georg Ostrovski, and Will Dabney. On the effect of auxiliary tasks on representation dynamics. In *International Conference on Artificial Intelligence and Statistics*, pp. 1–9. PMLR, 2021.

Xiaoteng Ma, Li Xia, Zhengyuan Zhou, Jun Yang, and Qianchuan Zhao. Dsac: Distributional soft actor critic for risk-sensitive reinforcement learning. *arXiv preprint arXiv:2004.14547*, 2020.

Yecheng Jason Ma, Dinesh Jayaraman, and Osbert Bastani. Conservative offline distributional reinforcement learning. *arXiv preprint arXiv:2107.06106*, 2021.

John Martin, Michal Lyskawinski, Xiaohu Li, and Brendan Englot. Stochastically dominant distributional reinforcement learning. In *International Conference on Machine Learning*, pp. 6745–6754. PMLR, 2020.

Borislav Mavrin, Shangtong Zhang, Hengshuai Yao, Linglong Kong, Kaiwen Wu, and Yaoliang Yu. Distributional reinforcement learning for efficient exploration. *International Conference on Machine Learning (ICML)*, 2019.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020.

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.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Parametric return density estimation for reinforcement learning. *arXiv preprint arXiv:1203.3497*, 2012.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019.

Thanh Tang Nguyen, Sunil Gupta, and Svetha Venkatesh. Distributional reinforcement learning with maximum mean discrepancy. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

Martin Riedmiller. Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In *European conference on machine learning*, pp. 317–328. Springer, 2005.

Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 29–37. PMLR, 2018.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. *International Conference on Machine Learning (ICML)*, 2019.

Ke Sun, Yi Liu, Yingnan Zhao, Hengshuai Yao, Shangling Jui, and Linglong Kong. Exploring the robustness of distributional reinforcement learning against noisy state observations. *arXiv preprint arXiv:2109.08776*, 2021.

Ke Sun, Yingnan Zhao, Yi Liu, Bei Jiang, and Linglong Kong. Distributional reinforcement learning via sinkhorn iterations. *arXiv preprint arXiv:2202.00769*, 2022.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An Introduction*. MIT press, 2018.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Ronald J Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. *Advances in neural information processing systems*, 32:6193–6202, 2019.

Fan Zhou, Jianing Wang, and Xingdong Feng. Non-crossing quantile regression for distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 2020.

A PROOF OF PROPOSITION 1

Proof. Firstly, we denote $Y = X - \mathbb{E}(X)$ and thus $F(\mathbb{E}\left[Z^\pi(s,a)\right]) = P(X \leq \mathbb{E}\left[Z^\pi(s,a)\right]) = 1 - P(Y > 0)$. Consider the bounded X case, we have the following inequalities:

$$\mathbb{1}_{Y>0} \geq \frac{cY + Y^2}{2c^2}, \text{and } \mathbb{1}_{Y>0} \leq \frac{cY - Y^2}{2c^2} + 1. \tag{13}$$

We take expectations on two sides and they arrive

$$P(Y > 0) \geq \frac{\mathbb{E}Y^2}{2c^2} = \frac{\sigma^2}{2c^2}, \text{and } P(Y > 0) \leq \frac{-\mathbb{E}Y^2}{2c^2} + 1 = 1 - \frac{\sigma^2}{2c^2}. \tag{14}$$

This indicates that $\frac{\sigma^2}{2c^2} \leq 1 - F(\mathbb{E}\left[Z^\pi(s,a)\right]) \leq 1 - \frac{\sigma^2}{2c^2}$. To put them together, we have the following results:

$$\begin{aligned}
\inf_{F_\mu^{s,a}} \|F - F^{s,a}\|_\infty &\equiv \inf_{F_\mu^{s,a}} \sup_x |F(x) - F^{s,a}(x)| \\
&\leq \sup_x |F(x) - F^{s,a}(x)| \, |_{F_\mu^{s,a} = F} \\
&= \max\{\sup_x (1-\epsilon)(1 - F(x)), \sup_x (1-\epsilon)F(x)\} \\
&= (1-\epsilon)\max\{1 - F(\mathbb{E}\left[Z^\pi(s,a)\right]), F(\mathbb{E}\left[Z^\pi(s,a)\right])\} \\
&\leq (1-\epsilon)(1 - \frac{\sigma^2}{2c^2}),
\end{aligned} \tag{15}$$

where the max operation is over two cases whether $x \geq \mathbb{E}\left[Z^\pi(s,a)\right]$ or not. The last inequality holds when we consider the bounded X case and it is also known that $\sigma^2 \leq c^2$. Therefore, in this bounded case, we can achieve a concise and uniform upper bound. ☐

B PROOF OF PROPOSITION 2

Proof. We firstly assume Z_θ is absolutely continuous and the supports of two distributions in KL divergence have a negligible intersection (Arjovsky & Bottou, 2017), under which the KL divergence is well-defined.

(1) Please refer to (Morimura et al., 2012) for the proof. Therefore, we have $D_{\text{KL}}^\infty(\mathfrak{T}^\pi Z_1, \mathfrak{T}^\pi Z_2) \leq D_{\text{KL}}^\infty(Z_1, Z_2)$, implying that $\mathfrak{T}^\pi$ is a non-expansive operator under D_{KL}^∞.

(2) By the definition of D_{KL}^∞, we have $\sup_{s,a} D_{\text{KL}}(Z_n(s,a), Z(s,a)) \to 0$ implies $D_{\text{KL}}(Z_n, Z) \to 0$. $D_{\text{KL}}(Z_n, Z) \to 0$ implies the total variation distance $\delta(Z_n, Z) \to 0$ according to a straightforward application of Pinsker's inequality

$$\begin{aligned}
\delta\left(Z_n, Z\right) &\leq \sqrt{\frac{1}{2} D_{\text{KL}}\left(Z_n, Z\right)} \to 0 \\
\delta\left(Z, Z_n\right) &\leq \sqrt{\frac{1}{2} D_{\text{KL}}\left(Z, Z_n\right)} \to 0
\end{aligned} \tag{16}$$

Based on Theorem 2 in WGAN (Arjovsky et al., 2017), $\delta(Z_n, Z) \to 0$ implies $W_p(Z_n, Z) \to 0$. This is trivial by recalling the fact that δ and W give the strong an weak topologies on the dual of $(C(\mathcal{X}), \|\cdot\|_\infty)$ when restricted to $\text{Prob}(\mathcal{X})$.

(3) The conclusion holds because the $\mathfrak{T}^\pi$ degenerates to $\mathcal{T}^\pi$ regardless of the metric d_p (Bellemare et al., 2017a). Specifically, due to the linearity of expectation, we obtain that

$$\|\mathbb{E}\mathfrak{T}^\pi Z_1 - \mathbb{E}\mathfrak{T}^\pi Z_2\|_\infty = \|\mathcal{T}^\pi \mathbb{E}Z_1 - \mathcal{T}^\pi \mathbb{E}Z_2\|_\infty \leq \gamma \|\mathbb{E}Z_1 - \mathbb{E}Z_2\|_\infty. \tag{17}$$

This implies that the expectation of Z under D_{KL} exponentially converges to the expectation of Z^*, i.e., γ-contraction. ☐

C PROOF OF PROPOSITION 3

Proof. Firstly, given a fixed $p(x)$ we know that minimizing $D_{\mathrm{KL}}(p, q_\theta)$ is equivalent to minimizing $\mathcal{H}(p, q)$ by following

$$\begin{aligned} D_{\mathrm{KL}}(p, q_\theta) &= \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q_\theta(x)} \,\mathrm{d}x \\ &= -\int_{-\infty}^{+\infty} p(x) \log q_\theta(x) \,\mathrm{d}x - \Big(-\int_{-\infty}^{+\infty} p(x) \log p(x) \,\mathrm{d}x\Big) \\ &= \mathcal{H}(p, q_\theta) - \mathcal{H}(p) \\ &\propto \mathcal{H}(p, q_\theta) \end{aligned} \tag{18}$$

Based on $\mathcal{H}(p, q_\theta)$, we use $p^{s_i', \pi_Z(s_i')}(x)$ to denote the target probability density function of the random variable $R(s_i, a_i) + \gamma Z_{\theta^*}^{k}\,(s_i', \pi_Z(s_i'))$. Then, we can derive the objective function within each Neural FZI as

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \mathcal{H}(p^{s_i', \pi_Z(s_i')}(x), q_\theta^{s_i, a_i}(x)) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(-\int_{-\infty}^{+\infty} p^{s_i', \pi_Z(s_i')}(x) \log q_\theta^{s_i, a_i}(x) \,\mathrm{d}x\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(-\int_{-\infty}^{+\infty} \left((1-\epsilon)\delta_{\{x=\mathbb{E}\left[Z^\pi(s_i', \pi_Z(s_i'))\right]\}}(x) + \epsilon\mu^{s_i', \pi_Z(s_i')}(x)\right) \log q_\theta^{s_i, a_i}(x) \,\mathrm{d}x\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left[(1-\epsilon)\left(-\int_{-\infty}^{+\infty} \delta_{\{x=\mathbb{E}\left[Z^\pi(s_i', \pi_Z(s_i'))\right]\}}(x) \log q_\theta^{s_i, a_i}(x) \,\mathrm{d}x\right) + \epsilon\mathcal{H}(\mu^{s_i', \pi_Z(s_i')}, q_\theta^{s_i, a_i})\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \left[(1-\epsilon)\mathcal{H}(\delta_{\{x=\mathbb{E}\left[Z^\pi(s_i', \pi_Z(s_i'))\right]\}}, q_\theta^{s_i, a_i}) + \epsilon\mathcal{H}(\mu^{s_i', \pi_Z(s_i')}, q_\theta^{s_i, a_i})\right] \\ &\propto \frac{1}{n}\sum_{i=1}^{n} \mathcal{H}(\delta_{\{x=\mathbb{E}\left[Z^\pi(s_i', \pi_Z(s_i'))\right]\}}, q_\theta^{s_i, a_i}) + \alpha\mathcal{H}(\mu^{s_i', \pi_Z(s_i')}, q_\theta^{s_i, a_i}), \text{ where } \alpha = \frac{\epsilon}{1-\epsilon} > 0 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(-\log q_\theta^{s_i, a_i}(\mathbb{E}\left[Z^\pi(s_i', \pi_Z(s_i'))\right]) + \alpha\mathcal{H}(\mu^{s_i', \pi_Z(s_i')}, q_\theta^{s_i, a_i})\right), \end{aligned} \tag{19}$$

where the last equality holds based on the fact that $\int_{-\infty}^{\infty} f(x)\delta_a(x)dx = f(a)$ when the Dirac function $\delta_a(x)$ is centered at a. $\square$

D PROOF OF CONVERGENCE OF SOFT DISTRIBUTIONAL POLICY ITERATION IN THEOREM 1

D.1 PROOF OF SOFT DISTRIBUTIONAL POLICY EVALUATION IN LEMMA 1

Proof. Firstly, we plug in $V(s_{t+1})$ into RHS of the iteration in Eq. 8, then we obtain

$$\begin{aligned} &\mathcal{T}_d^\pi Q\left(s_t, a_t\right) \\ &= r\left(s_t, a_t\right) + \gamma \mathbb{E}_{s_{t+1} \sim \rho^\pi}\left[V\left(s_{t+1}\right)\right] \\ &= r\left(s_t, a_t\right) + \gamma f(\mathcal{H}\left(\mu^{s_t, a_t}, q_\theta^{s_t, a_t}\right)) + \gamma \mathbb{E}_{(s_{t+1}, a_{t+1}) \sim \rho^\pi}\left[Q\left(s_{t+1}, a_{t+1}\right)\right] \\ &\triangleq r_\pi\left(s_t, a_t\right) + \gamma \mathbb{E}_{(s_{t+1}, a_{t+1}) \sim \rho^\pi}\left[Q\left(s_{t+1}, a_{t+1}\right)\right], \end{aligned} \tag{20}$$

where $r_\pi\left(s_t, a_t\right) \triangleq r\left(s_t, a_t\right) + \gamma f(\mathcal{H}\left(\mu^{s_t, a_t}, q_\theta^{s_t, a_t}\right))$ is the entropy augmented reward we redefine. Applying the standard convergence results for policy evaluation (Sutton & Barto, 2018), we can attain that this Bellman updating under $\mathcal{T}_{sd}^\pi$ is convergent under the assumption of $|\mathcal{A}| < \infty$ and bounded entropy augmented rewards r_π. $\square$

D.2 POLICY IMPROVEMENT WITH PROOF

Lemma 2. *(Distribution-Entropy-Regularized Policy Improvement) Let* $\pi \in \Pi$ *and a new policy* π_{new} *be updated via the policy improvement step in Eq. 11. Then* $Q^{\pi_{new}}\left(s_t, a_t\right) \geq Q^{\pi_{old}}\left(s_t, a_t\right)$ *for all* $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$ *with* $|\mathcal{A}| \leq \infty$.

The policy improvement in Lemma 2 implies that $\mathbb{E}_{a_t \sim \pi_{\text{new}}}\left[Q^{\pi_{\text{old}}}(s_t, a_t)\right] \geq \mathbb{E}_{a_t \sim \pi_{\text{old}}}\left[Q^{\pi_{\text{old}}}(s_t, a_t)\right]$, we consider the Bellman equation via the distribution-entropy-regularized Bellman operator $\mathcal{T}_{sd}^{\pi}$:

$$
\begin{aligned}
Q^{\pi_{\text{old}}}\left(s_t, a_t\right) & \triangleq r\left(s_t, a_t\right)+\gamma \mathbb{E}_{s_{t+1} \sim \rho}\left[V^{\pi_{\text{old}}}\left(s_{t+1}\right)\right] \\
& =r\left(s_t, a_t\right)+\gamma f(\mathcal{H}\left(\mu^{s_t, a_t}, q_\theta^{s_t, a_t}\right))+\gamma \mathbb{E}_{\left(s_{t+1}, a_{t+1}\right) \sim \rho^{\pi_{\text{old}}}}\left[Q^{\pi_{\text{old}}}\left(s_{t+1}, a_{t+1}\right)\right] \\
& \leq r\left(s_t, a_t\right)+\gamma f(\mathcal{H}\left(\mu^{s_t, a_t}, q_\theta^{s_t, a_t}\right))+\gamma \mathbb{E}_{\left(s_{t+1}, a_{t+1}\right) \sim \rho^{\pi_{\text{new}}}}\left[Q^{\pi_{\text{old}}}\left(s_{t+1}, a_{t+1}\right)\right] \\
& =r_{\pi^{\text{new}}}\left(s_t, a_t\right)+\gamma \mathbb{E}_{\left(s_{t+1}, a_{t+1}\right) \sim \rho^{\pi_{\text{new}}}}\left[Q^{\pi_{\text{old}}}\left(s_{t+1}, a_{t+1}\right)\right] \\
& \vdots \\
& \leq Q^{\pi_{\text{new}}}\left(s_{t+1}, a_{t+1}\right),
\end{aligned}
\tag{21}
$$

where we have repeated expanded $Q^{\pi_{\text{old}}}$ on the RHS by applying the distribution-entropy-regularized distributional Bellman operator. Convergence to $Q^{\pi_{\text{new}}}$ follows from Lemma 1.

D.3 PROOF OF SOFT DISTRIBUTIONAL POLICY ITERATION IN THEOREM 1

The proof is similar to soft policy iteration (Haarnoja et al., 2018). For the completeness, we provide the proof here. By Lemma 2, as the number of iteration increases, the sequence Q^{π_i} at i-th iteration

Table 1: Hyper-parameters Sheet.

Hyperparameter	Value
Shared	
Policy network learning rate	3e-4
(Quantile) Value network learning rate	3e-4
Optimization	Adam
Discount factor	0.99
Target smoothing	5e-3
Batch size	256
Replay buffer size	1e6
Minimum steps before training	1e4
DSAC with C51	
Number of Atoms (N)	51
DSAC with IQN	
Number of quantile fractions (N)	32
Quantile fraction embedding size	64
Huber regression threshold	1

Hyperparameter	Temperature Parameter β	Max episode lenght
Walker2d-v2	0.2	1000
Swimmer-v2	0.2	1000
Reacher-v2	0.2	1000
Ant-v2	0.2	1000
HalfCheetah-v2	0.2	1000
Humanoid-v2	0.05	1000
HumanoidStandup-v2	0.05	1000
BipedalWalkerHardcore-v2	0.002	2000

is monotonically increasing. Since we assume the risk-sensitive entropy is bounded by M, the Q^π is thus bounded as the rewards are bounded. Hence, the sequence will converge to some π^*. Further, we prove that π^* is in fact optimal. At the convergence point, for all $\pi \in \Pi$, it must be case that:

$$\mathbb{E}_{a_t \sim \pi^*}\left[Q^{\pi_{\text{old}}}\left(s_t, a_t\right)\right] \geq \mathbb{E}_{a_t \sim \pi}\left[Q^{\pi_{\text{old}}}\left(s_t, a_t\right)\right].$$

According to the proof in Lemma 2, we can attain $Q^{\pi^*}(s_t, a_t) > Q^\pi(s_t, a_t)$ for (s_t, a_t). That is to say, the "corrected" value function of any other policy in Π is lower than the converged policy, indicating that π^* is optimal.

E IMPLEMENTATION DETAILS

Our implementation is directly adapted from the source code in (Ma et al., 2020).

For Distributional SAC with C51, we use 51 atoms similar to the C51 (Bellemare et al., 2017a). For distributional SAC with quantile regression, instead of using fixed quantiles in QR-DQN (Dabney et al., 2018b), we leverage the quantile fraction generation based on IQN (Dabney et al., 2018a) that uniformly samples quantile fractions in order to approximate the full quantile function. In particular, we fix the number of quantile fractions as N and keep them in an ascending order. Besides, we adapt the sampling as $\tau_0 = 0, \tau_i = \epsilon_i / \sum_{i=0}^{N-1}$, where $\epsilon_i \in U[0, 1], i = 1, ..., N$.

E.1 HYPER-PARAMETERS AND NETWORK STRUCTURE.

We adopt the same hyper-parameters, which is listed in Table 1 and network structure as in the original distributional SAC paper (Ma et al., 2020).

F DERAC ALGORITHM

Algorithm 1 Distribution-Entropy-Regularized Actor Critic (DERAC) Algorithm

```
Initialize two value networks $q_\theta$, $q_{\theta^*}$, and policy network $\pi_\phi$.
for each iteration do
    for each environment step do
        $a_t \sim \pi_\phi(a_t|s_t)$.
        $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$.
        $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r(s_t, a_t), s_{t+1})\}$
    end for
    for each gradient step do
        $\theta \leftarrow \theta - \lambda_q \nabla_\theta \hat{J}_q(\theta)$
        $\phi \leftarrow \phi + \lambda_\pi \nabla_\phi \hat{J}_\pi(\phi)$.
        $\theta^* \leftarrow \tau\theta + (1-\tau)\theta^*$
    end for
end for
```

G EXPERIMENTS

Figure 5 suggests that C51 with cross entropy loss behaves similarly to the vanilla C51 equipped with KL divergence.

Figure 6 shows that DERAC with different λ in Eq. 12 may behave differently on the different environment. Learning curves of DERAC with an increasing λ will tend to DSAC (C51), e.g., Bipedalwalkerhardcore, where DERAC with $\lambda = 1$ in the green line tends to DSAC (C51) in the blue line. However, DERAC with a small λ is likely to outperform DSAC (C51) by only leverage the expectation effect of value distribution, e.g., on Bipedalwalkerhardcore, where DERAC with $\lambda = 0, 0.5$ bypass DERAC with $\lambda = 1.0$.

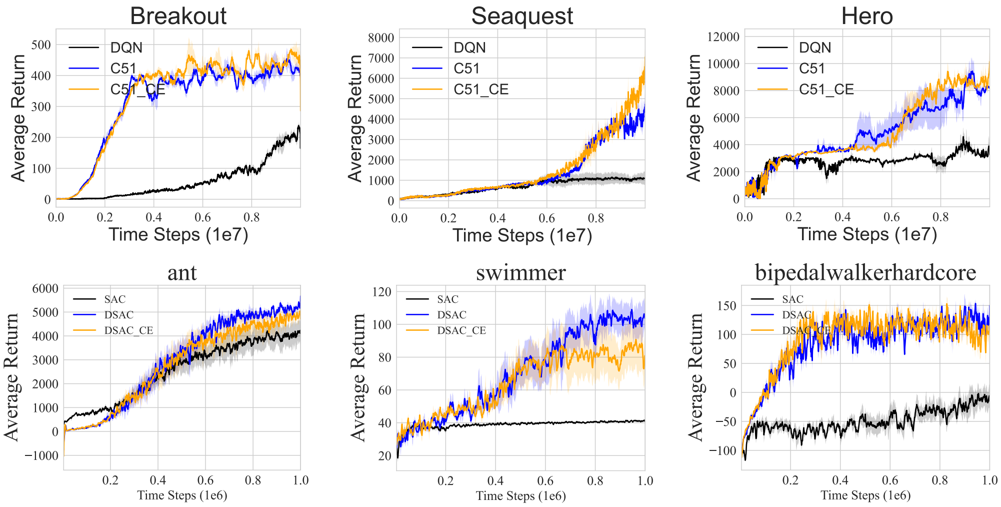

Figure 5: (**First row**) Learning curves of C51 under cross entropy loss on Atari games over 3 seeds. (**Second row**) Learning curves of DSAC with C51 under cross entropy loss on MuJoCo environments over 5 seeds.

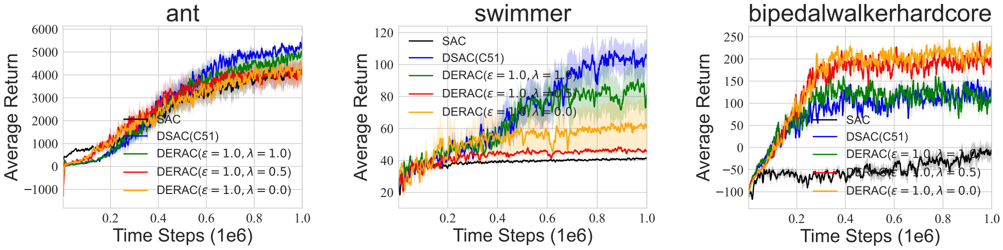

Figure 6: Learning curves of DERAC algorithms across different λ on three MuJoCo environments over 5 seeds.