Intrinsic Benefits of Categorical Distributional Loss: Uncertainty-aware Exploration in Reinforcement Learning towards Higher Moment Regularisations

Ke Sun[1], **Yingnan Zhao**[2], **Enze Shi**[1], **Yafei Wang**[1], **Xiaodong Yan**[3], **Bei Jiang**[1], **Linglong Kong**[1]*
[1]University of Alberta, Canada
[2]Harbin Engineering University, China
[3]Shangdong University, China
{ksun6,eshi,yafei2,bei1,lkong}@ualberta.ca
zhaoyingnan@hrbeu.edu.cn
yanxiaodong@sdu.edu.cn

Abstract

The remarkable empirical performance of distributional reinforcement learning (RL) has garnered increasing attention to understanding its theoretical advantages over classical RL. By decomposing the categorical distributional loss commonly employed in distributional RL, we find that the potential superiority of distributional RL can be attributed to a derived distribution-matching entropy regularization that captures higher moment knowledge. This less-studied entropy regularization aims to capture additional knowledge of return distribution beyond only its expectation, contributing to an augmented reward signal in policy optimization. In contrast to the vanilla entropy regularization in MaxEnt RL, which explicitly encourages exploration by promoting diverse actions, the novel entropy regularization derived from categorical distributional loss implicitly updates policies to align the learned policy with (estimated) environmental uncertainty. Finally, extensive experiments verify the significance of this uncertainty-aware regularization from distributional RL on the empirical benefits over classical RL. Our study offers an innovative exploration perspective to explain the intrinsic benefits of distributional learning in RL.

1 Introduction

The fundamental characteristics of classical reinforcement learning (RL) [49], such as Q-learning [53], rely on estimating the expectation of discounted cumulative rewards that an agent observes while interacting with the environment. In contrast to the expectation-based RL, a novel branch of algorithms, termed *distributional RL*, seeks to estimate the entire distribution of total returns and has achieved state-of-the-art performance across a diverse array of environments [2, 8, 7, 59, 35, 54, 48, 44]. Meanwhile, discussions of distributional RL have increasingly extended into a broader range of fields, such as risk-sensitive control [7, 25, 5], offline learning [29, 58], policy exploration [30, 41, 6, 20], robustness [47, 45, 43], optimization [43, 22, 46], statistical inference [60], multivariate rewards [61, 57], and continuous-time setting [56].

Motivation: Understanding the Benefits of Employing (Categorical) Distributional Loss in RL. Despite the impressive empirical success of distributional RL algorithms, our comprehension of their advantages over classical RL remains incomplete, especially for the general function approximation setting and practical implementations. Early work [27] demonstrated that in many realizations of

*Corresponding author

First Exploration in AI Today Workshop at ICML (EXAIT at ICML 2025).

tabular and linear approximation settings, distributional RL behaves similarly to classic RL, suggesting that its benefits are mainly realized in the non-linear approximation setting. Although their findings offer profound insights, their analysis, based on a coupled update method, overlooks several factors, such as the optimization effect under various losses. The statistical benefits of quantile temporal difference (QTD), employed in quantile distributional RL, e.g., QR-DQN [8], were highlighted in [43, 42], which posited that the robust estimation of QTD fosters the benefits in stochastic environments. The foundational theoretical aspects of Categorical Distributional RL (CDRL), e.g., C51 [2], were first discussed in [40]; however, explaining the advantages of categorical distributional learning remains under-explored. Recent studies [52, 51] elucidate the benefits of distributional RL by introducing the small-loss and second-order PAC bounds, revealing the enhanced sample efficiency, particularly in specific cases with small achievable costs. Yet, their findings are not directly based on practical distributional RL algorithms, such as C51 or QR-DQN. Therefore, it is imperative to close this gap between understanding their theoretical advantages and practical deployment in complex environments for distributional RL algorithms. More related work is provided in Appendix A.

Contributions. In this study, we interpret the potential superiority of distributional learning in RL over classical RL, specifically focusing on CDRL, the pioneering family within distributional RL. We examine the benefits through the lens of a regularized exploration effect, offering a distinct perspective relative to existing literature. Our investigation begins by decomposing the categorical distributional loss into a mean-related term and a distribution-matching regularization term, facilitated by our proposed return density decomposition technique. The resulting regularization acts as an augmented reward in the actor critic framework, encouraging policies to explore states whose *current return distribution estimates lag far behind the (estimated) environmental uncertainty in the target return*. This derived regularization from the categorical distributional loss in CDRL promotes an uncertainty-aware exploration effect, which diverges from the exploration for diverse actions commonly used in MaxEnt RL [55, 13, 14]. We also provide the convergence foundations when leveraging the decomposed uncertain-aware regularization in the actor critic. Empirical evidence underscores the pivotal role of the uncertainty-aware entropy regularization in the empirical success of adopting categorical distributional loss in RL over classical RL on both Atari games and MuJoCo tasks. We further elucidate the distinct roles that the uncertainty-aware entropy in distributional RL and the vanilla entropy in MaxEnt RL play by exploring their mutual impacts on learning performance. This opens new avenues for future research in this domain. Our contributions are summarized as follows:

1. By applying a return density decomposition on the categorical distributional loss, we derive a distribution-matching regularization. This regularization promotes uncertainty-aware exploration, interpreting the benefits of categorical distributional learning in RL.
2. We extend the benefit interpretation of the categorical distributional loss to policy gradient methods. We compare the different exploration effects of our decomposed uncertainty-aware regularization from distributional RL and the vanilla entropy regularization in MaxEnt RL.
3. Empirically, we verify the uncertainty-aware regularization effect on the performance improvement of distributional RL and investigate the mutual impacts of two regularizations in learning.

2 Preliminaries

Markov Decision Process (MDP) and Classical RL. An environment is modeled via an Markov Decision Process $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P, \gamma)$, with a set of states $\mathcal{S}$ and actions $\mathcal{A}$, the bounded reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathcal{P}([R_{\min}, R_{\max}])$, the transition kernel $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$, and a discounted factor $\gamma \in [0, 1]$. We denote the reward the agent receives at time t as $r_t \sim \mathcal{R}(s_t, a_t)$. Given a policy π, the key quantity of interest is the return Z^π, which is the total cumulative rewards over the course of a trajectory defined by $Z^\pi(s, a) = \sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a$. Classical RL focuses on estimating the expectation of the return, i.e., $Q^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t=0}^{+\infty} \gamma^t r_t | s_0 = s, a_0 = a \right]$. We also define Bellman evaluation operator $\mathcal{T}^\pi Q(s, a) = \mathbb{E}[\mathcal{R}(s, a)] + \gamma \mathbb{E}_{s' \sim P, a' \sim \pi} \left[Q\left(s', a'\right) \right]$, and Bellman optimality operator $\mathcal{T}^{\text{opt}} Q(s, a) = \mathbb{E}[\mathcal{R}(s, a)] + \gamma \max_{a'} \mathbb{E}_{s' \sim P} \left[Q\left(s', a'\right) \right]$.

Distributional RL and CDRL. Instead of only learning the expectation in classical RL, distributional RL models the full distribution of the return Z^π. The return distribution $\eta^\pi : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathbb{R})$ is defined as $\eta^\pi(s, a) = \mathcal{D}(Z^\pi(s, a))$, where $\mathcal{D}$ extracts the distribution of a random variable. $\eta^\pi(s, a)$ is updated via the distributional Bellman operator $\mathfrak{T}^\pi$, defined by $\mathfrak{T}^\pi Z(s, a) \stackrel{D}{=} \mathcal{R}(s, a) +$

$\gamma Z\left(s', a'\right)$, where $\stackrel{D}{=}$ implies that random variables of both sides are equal in distribution. Categorical Distributional RL (CDRL) is the first successful distributional RL family that approximates the return distribution by a discrete categorical distribution $\widehat{\eta}^{\pi} = \sum_{i=1}^{N} p_i \delta_{z_i}$, where $\{z_i\}_{i=1}^{N}$ is a set of fixed supports and $\{p_i\}_{i=1}^{N}$ are learnable probabilities. The leverage of a heuristic projection operator $\Pi_{\mathcal{C}}$ (see Appendix B for more details) and the Kullback–Leibler (KL) divergence guarantee the theoretical convergence of CDRL under Cramér distance or Wasserstein distance in the tabular setting [40].

3 Regularization Benefits in Value-based Distribution RL

3.1 Distributional RL: Neural FZI

Classical RL: Neural Fitted Q-Iteration (Neural FQI). Neural FQI [9, 39] offers a statistical explanation of DQN [32], capturing its key features, including experience replay and the target network Q_{θ^*}. We update a parameterized Q_θ in each iteration k of an iterative regression:

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

where the target $y_i^k = r(s_i, a_i) + \gamma \max_{a \in \mathcal{A}} Q_{\theta^*}^k\left(s_i', a\right)$ is fixed within every T_{target} steps to update target network Q_{θ^*} by letting $Q_{\theta^*}^k = Q_\theta^k$. The experience buffer induces independent samples $\{(s_i, a_i, r_i, s_i')\}_{i \in [n]}$. If $\{Q_\theta : \theta \in \Theta\}$ is sufficiently large such that it contains $\mathcal{T}^{\text{opt}} Q_{\theta^*}^k$, i.e., the realizable assumption in learning theory [33], Neural FQI has the solution $Q_\theta^{k+1} = \mathcal{T}^{\text{opt}} Q_{\theta^*}^k$, which is exactly the updating rule under Bellman optimality operator [9].

Distributional RL: Neural Fitted Z-Iteration (Neural FZI). Analogous to Neural FQI, we simplify value-based distributional RL algorithms with the parameterized Z_θ as Neural FZI:

$$Z_\theta^{k+1} = \underset{Z_\theta}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} d_p(Y_i^k, Z_\theta\left(s_i, a_i\right)), \tag{2}$$

where we denote the target return as $Y_i^k = \mathcal{R}(s_i, a_i) + \gamma Z_{\theta^*}^k\left(s_i', \pi_Z(s_i')\right)$ with the policy π_Z following the greedy rule $\pi_Z(s_i') = \text{argmax}_{a'} \mathbb{E}\left[Z_{\theta^*}^k(s_i', a')\right]$. The target Y_i^k is fixed within every T_{target} steps to update target network Z_{θ^*}. d_p is a distribution divergence between two distributions. While our analysis is not intended to involve properties of deep neural networks, we interpret distributional RL as Neural FZI, as it is by far the closest to the practical algorithms.

3.2 Distributional RL: Entropy-regularized Neural FQI

As mentioned previously in preliminary knowledge in Section 2, CDRL employs neural networks to learn the probabilities $\{p_i\}_{i=1}^{N}$ in a discrete categorical distribution to represent Z_θ, and choose KL divergence as d_p in Eq. 2 of Neural FZI. We next decompose the KL-based distributional loss d_p in CDRL by utilizing an equivalent histogram density estimator $\widehat{p}$ in representing Z_θ.

Return Density Decomposition. To characterize the impact of additional knowledge from the return distribution beyond its expectation, we use a variant of *gross error model* from robust statistics [18], which was also similarly applied to analyze Label Smoothing [34] and Knowledge Distillation [17]. Akin to the categorical parameterization in CDRL, we utilize a histogram function estimator $\widehat{p}^{s,a}(x)$ with N bins to approximate an arbitrary continuous density $p^{s,a}(x)$ of $Z^\pi(s, a)$, given a state s and action a. In contrast to categorical parameterization defined on a set of fixed supports, the histogram estimator operates over a continuous interval, enabling more nuanced analysis within continuous functions. Given a fixed set of supports $l_0 \leq l_1 \leq ... \leq l_N$ with the equal bin size as Δ, each bin is thus denoted as $\Delta_i = [l_{i-1}, l_i)$, $i = 1, ..., N-1$ with $\Delta_N = [l_{N-1}, l_N]$. As such, the histogram density estimator is formulated by $\widehat{p}^{s,a}(x) = \sum_{i=1}^{N} p_i \mathbb{1}(x \in \Delta_i)/\Delta$ with p_i as the coefficient in the i-th bin Δ_i. Denote

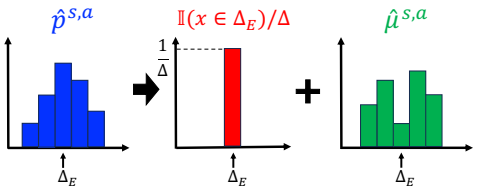

Figure 1: Return Density Decomposition on Histograms.

Δ_E as the interval that $\mathbb{E}[Z^\pi(s,a)]$ falls into, i.e., $\mathbb{E}[Z^\pi(s,a)] \in \Delta_E$. Putting all together, we apply an action-state return density decomposition over the histogram density estimator $\widehat{p}^{s,a}$:

$$\widehat{p}^{s,a}(x) = (1-\epsilon)\mathbb{1}(x \in \Delta_E)/\Delta + \epsilon\widehat{\mu}^{s,a}(x), \tag{3}$$

where $\widehat{p}^{s,a}$ is decomposed into a single-bin histogram $\mathbb{1}(x \in \Delta_E)/\Delta$ with all mass on Δ_E and an **induced** histogram density function $\widehat{\mu}^{s,a}$ evaluated by $\widehat{\mu}^{s,a}(x) = \sum_{i=1}^N p_i^\mu \mathbb{1}(x \in \Delta_i)/\Delta$ with p_i^μ as the coefficient of the i-th bin Δ_i. ϵ is a hyper-parameter pre-specified before the decomposition, controlling the proportion between $\mathbb{1}(x \in \Delta_E)/\Delta$ and $\widehat{\mu}^{s,a}(x)$. See Figure 1 for the illustration of the decomposition. More specifically, the induced histogram density function $\widehat{\mu}^{s,a}$ in the second term of Eq. 3 represents the difference between the full histogram function $\widehat{p}^{s,a}$ and a single-bin histogram $\mathbb{1}(x \in \Delta_E)/\Delta$, where $\mathbb{1}(x \in \Delta_E)/\Delta$ only captures the mean. This difference indicates that $\widehat{\mu}^{s,a}$ *captures the additional distribution information of* $Z^\pi(s,a)$ *beyond its expectation* $\mathbb{E}[Z^\pi(s,a)]$*, incorporating higher-moments information*. This reflects the influence of using a full distribution on the performance of distributional RL. The additional leverage of $\widehat{\mu}^{s,a}$ in the distributional loss explains the behavior differences between classical and distribution RL algorithms. We next demonstrate that $\widehat{\mu}^{s,a}$ is a valid probability density under certain ϵ in Proposition 1.

Proposition 1. *(Decomposition Validity) Denote* $\widehat{p}^{s,a}(x \in \Delta_E) = p_E \frac{\mathbb{1}(x \in \Delta_E)}{\Delta}$*, where* p_E *is the coefficient on the bin* Δ_E*.* $\widehat{\mu}^{s,a}(x) = \sum_{i=1}^N p_i^\mu \mathbb{1}(x \in \Delta_i)/\Delta$ *is a valid density if and only if* $\epsilon \geq 1 - p_E$*.*

The proof can be found in Appendix C. Proposition 1 demonstrates that the return density decomposition is valid when the hyper-parameter ϵ is well specified as $\epsilon \geq 1 - p_E$. Under this condition, our analysis maintains the standard categorical distributional learning in distributional RL.

Distributional RL: Entropy-regularized Neural FQI. We apply the decomposition in Eq. 3 on the histogram density function, denoted as $\widehat{p}^{s_i',\pi_Z(s_i')}$, of the target return $Y_i^k = \mathcal{R}(s_i,a_i) + \gamma Z_{\theta^*}^k(s_i', \pi_Z(s_i'))$ in Eq. 2 of Neural FZI. Consequently, we have $\widehat{p}^{s_i',\pi_Z(s_i')}(x) = (1-\epsilon)\mathbb{1}(x \in \Delta_E^i)/\Delta + \epsilon\widehat{\mu}^{s_i',\pi_Z(s_i')}(x)$, where Δ_E^i represents the interval that the expectation of the target return Y_i^k falls into, i.e., $\mathbb{E}\left[Y_i^k\right] \in \Delta_E^i$, and $\widehat{\mu}^{s_i',\pi_Z(s_i')}$ is the induced histogram density function, similar to the role of $\widehat{\mu}^{s,a}$ in Eq. 3. Let $\mathcal{H}(U,V)$ be the cross-entropy between two probability measures U and V, i.e., $\mathcal{H}(U,V) = -\int_{x\in\mathcal{X}} U(x)\log V(x)\,\mathrm{d}x$. Immediately, we can derive the following entropy-regularized loss function form of Neural FZI for distributional RL in Proposition 2. The proof is provided in Appendix D.

Proposition 2. *(Decomposed Neural FZI) Denote* $q_\theta^{s,a}$ *as the histogram density estimator of* $Z_\theta^k(s,a)$ *in Neural FZI. Based on the decomposition in Eq. 3 and the KL divergence as* d_p*, the Neural FZI process in Eq. 2 is simplified as*

$$Z_\theta^{k+1} = \operatorname*{argmin}_{q_\theta} \frac{1}{n}\sum_{i=1}^n [\underbrace{-\log q_\theta^{s_i,a_i}(\Delta_E^i)}_{\textit{Mean-Related Term}} + \underbrace{\alpha\mathcal{H}(\widehat{\mu}^{s_i',\pi_Z(s_i')}, q_\theta^{s_i,a_i})}_{\textit{Regularization Term}}], \tag{4}$$

where $\alpha = \varepsilon/(1-\varepsilon) > 0$ *and the mean-related term is negative log-likelihood centered on* Δ_E^i*.*

Connection between Neural FQI and FZI. A crucial bridge between classical RL and distributional RL is established in Proposition 3, where we demonstrate that minimizing the mean-related term in Eq. 4 of Neural FZI is asymptotically equivalent to minimizing Neural FQI in terms of the minimizers as $\Delta \to 0$. As such, with this equivalence in the objective function, the remaining regularization term $\alpha\mathcal{H}(\widehat{\mu}^{s_i',\pi_Z(s_i')}, q_\theta^{s_i,a_i})$ in Eq. 4 thus interprets the potential benefits of CDRL over classical RL. For the uniformity of notation, we still use s,a in the following analysis instead of s_i,a_i.

Proposition 3. *(Equivalence between* ***the Mean-Related term*** *in Decomposed Neural FZI and Neural FQI) In Eq. 4, assume the function class* $\{Z_\theta : \theta \in \Theta\}$ *is sufficiently large such that it contains the target* $\{Y_i^k\}_{i=1}^n$ *for all* k*, when* $\Delta \to 0$*, minimizing* ***the mean-related term*** *in Eq. 4 implies*

$$\mathbb{P}(Z_\theta^{k+1}(s,a) = \mathcal{T}^{opt}Q_{\theta^*}^k(s,a)) = 1, \tag{5}$$

where $\mathcal{T}^{opt}Q_{\theta^*}^k(s,a)$ *is the scalar-valued target in the k-th phase of Neural FQI.*

Proposition 3 demonstrates that as $\Delta \to 0$, the random variable $Z_\theta^{k+1}(s,a)$ with the limiting distribution in Neural FZI (distributional RL) will *degrade* to a constant $\mathcal{T}^{\text{opt}}Q_{\theta^*}^k(s,a)$, the minimizer

(scalar-valued target) in Neural FQI (classical RL). That being said, *minimizing the mean-related term in Neural FZI is asymptotically equivalent to minimizing Neural FQI with the same limiting minimizer*. A formal proof for convergence in distribution with the convergence rate $o(\Delta)$ is given in Appendix E. The realizable assumption that $\{Z_\theta : \theta \in \Theta\}$ is sufficiently large such that it contains $\{Y_i^k\}_{i=1}^n$ implies good in-distribution generalization performance in each phase of Neural FZI, which is also adopted in [58]. This connection is also consistent with the mean-preserving property of distributional RL in the tabular setting [40], but we extend this conclusion to the arbitrary function approximation with a histogram density estimator. Proposition 3 especially focuses on the asymptotic property of the mean-related term, which is different from existing convergence results based on the entire categorical distribution [40, 3]. Given the connection between optimizing the mean-related term of Neural FZI with Neural FQI in Proposition 3, we can leverage the regularization term $\alpha\mathcal{H}(\widehat{\mu}^{s_i',\pi_Z(s_i')}, q_\theta^{s_i,a_i})$ to explain the behavior difference between CDRL and classical RL, as analyzed later.

3.3 Uncertainty-aware Regularized Exploration

Regularization Effect. It turns out that minimizing the regularization term $\alpha\mathcal{H}(\widehat{\mu}^{s_i',\pi_Z(s_i')}, q_\theta^{s_i,a_i})$ in Neural FZI pushes $q_\theta^{s,a}$ for the current return density estimator to catch up with the target return density function of $\widehat{\mu}^{s_i',\pi_Z(s_i')}$. Importantly, $\widehat{\mu}^{s_i',\pi_Z(s_i')}$ encompasses the uncertainty of the entire return distribution in the learning course beyond only its expectation, given that $\widehat{\mu}^{s_i',\pi_Z(s_i')}$ is the induced histogram density after applying the return density decomposition in Eq. 3. Since it is a prevalent notion that distributional RL can significantly reduce intrinsic uncertainty of the environment [30, 7], the derived distribution-matching regularization term $\alpha\mathcal{H}(\widehat{\mu}^{s_i',\pi_Z(s_i')}, q_\theta^{s_i,a_i})$ helps to capture more uncertainty of the environment by modeling higher moments of the whole return distribution beyond the expectation. In Section 4, we further demonstrate that this derived regularization contributes to an *uncertainty-aware regularized exploration* effect in the policy optimization or actor critic.

Remark: Approximation and Calculation of $\widehat{\mu}^{s',\pi_Z(s')}$. In practical distributional RL algorithms, we typically use temporal-difference (TD) learning to attain the target probability density estimate $\widehat{\pi}^{s',\pi_Z(s')}$. Then we evaluate $\widehat{\mu}^{s',\pi_Z(s')}$ based on the decomposition in Eq. 3, provided $\mathbb{E}[Z(s,a)]$ exists and $\epsilon \geq 1 - p_E$ in Proposition 1. The approximation error of $\widehat{\mu}^{s',\pi_Z(s')}$ is fundamentally determined by the TD learning nature. A desirable approximation of $\widehat{\mu}^{s',\pi_Z(s')}$ intuitively leads to performance improvement in distributional RL.

4 Regularization Benefits in Actor Critic

4.1 Connection with MaxEnt RL

Explicit Entropy Regularization in MaxEnt RL. MaxEnt RL *explicitly* encourages exploration by optimizing for policies to reach states with higher entropy in the future:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\rho_\pi}\left[r\left(\mathbf{s}_t,\mathbf{a}_t\right) + \beta\mathcal{H}(\pi(\cdot|\mathbf{s}_t))\right], \tag{6}$$

where $\mathcal{H}\left(\pi_\theta\left(\cdot|\mathbf{s}_t\right)\right) = -\sum_a \pi_\theta\left(a|\mathbf{s}_t\right)\log\pi_\theta\left(a|\mathbf{s}_t\right)$ 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. 6.

Implicit Entropy Regularization in Distributional RL. For a direct comparison with MaxEnt RL, it is required to specifically analyze the impact of the regularization term in Eq. 4. Therefore, we directly incorporate the distribution-matching regularization of distributional RL in Eq. 4 into the Actor Critic (AC) framework, enabling us to consider a new soft Q-value. The new Q function can be computed iteratively by applying a modified Bellman operator denoted as $\mathcal{T}_d^\pi$, called *Distribution-Entropy-Regularized Bellman Operator*. Given a fixed q_θ, $\mathcal{T}_d^\pi$ is defined as

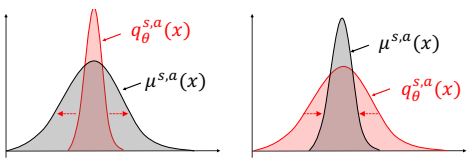

Figure 2: $q_\theta^{s,a}$ is optimized to disperse (left) or concentrate (right) to align with the uncertainty of target return distributions.

$$\mathcal{T}_d^\pi Q\left(\mathbf{s}_t,\mathbf{a}_t\right) \triangleq r\left(\mathbf{s}_t,\mathbf{a}_t\right) + \gamma\mathbb{E}_{\mathbf{s}_{t+1}\sim P(\cdot|\mathbf{s}_t,\mathbf{a}_t)}\left[V\left(\mathbf{s}_{t+1}\right)\right], \tag{7}$$

where a new soft value function $V(\mathbf{s}_t)$ is defined by

$$V(\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t \sim \pi}\left[Q(\mathbf{s}_t, \mathbf{a}_t) + f(\mathcal{H}(\mu^{\mathbf{s}_t,\mathbf{a}_t}, q_\theta^{\mathbf{s}_t,\mathbf{a}_t}))\right], \tag{8}$$

where f is a continuous increasing function over the cross-entropy $\mathcal{H}$. $\mu^{\mathbf{s}_t,\mathbf{a}_t}$ is the induced true target return histogram density function via the decomposition in Eq. 3, which excludes its expectation. Note that $\mu^{\mathbf{s}_t,\mathbf{a}_t}$ can be approximated via bootstrap TD estimate $\widehat{\mu}^{\mathbf{s}_{t+1},\pi_Z(\mathbf{s}_{t+1})}$ similar to Eq. 4. In this specific tabular setting regarding $\mathbf{s}_t, \mathbf{a}_t$, we particularly use $q_\theta^{\mathbf{s}_t,\mathbf{a}_t}$ to approximate the true density function of $Z(\mathbf{s}_t, \mathbf{a}_t)$. The f transformation over the cross-entropy $\mathcal{H}$ between $\mu^{\mathbf{s}_t,\mathbf{a}_t}$ and $q_\theta^{\mathbf{s}_t,\mathbf{a}_t}(x)$ serves as the uncertainty-aware entropy regularization that we implicitly derive from value-based distributional RL in Section 3.2. By optimizing q_θ that is involved in the value-based critic component in actor critic, this regularization reduces the mismatch between the target return distribution and current estimate, aligning with the regularization effect analyzed in Section 3.3. As illustrated in Figure 2, $q_\theta^{s,a}$ is optimized to **catch up with** the uncertainty involved in the target return distribution of $\mu^{s,a}$, iteratively expanding the agent's knowledge about the environment uncertainty to contribute to more informative decisions. Next, we elaborate on its additional impact on policy learning in the actor critic compared to MaxEnt RL.

Reward Augmentation for Policy Learning. As opposed to the vanilla entropy regularization in MaxEnt RL that explicitly encourages the policy to explore, our derived regularization term in the distributional loss of RL plays the role of reward augmentation for policy learning. Compared with classical RL, the augmented reward from the distributional loss incorporates additional knowledge of the return distribution in the learning process. As we will show later, *the augmented reward encourages policies to reach states $\mathbf{s}_t$ with actions $\mathbf{a}_t \sim \pi(\cdot|\mathbf{s}_t)$, whose current action-state return distribution $q_\theta^{\mathbf{s}_t,\mathbf{a}_t}$* ***lags far behind*** *the (estimated) environmental uncertainty from the target returns*.

For a detailed comparison with MaxEnt RL, we now focus on the properties of our decomposed distribution-matching regularization in the actor critic. In Lemma 1, we demonstrate that Distribution-Entropy-Regularized Bellman operator $\mathcal{T}_d^\pi$ inherits the convergence property in the policy evaluation phase with a cumulative augmented reward function as the new objective function $J'(\pi)$.

Lemma 1. *(Distribution-Entropy-Regularized Policy Evaluation) Consider the distribution-entropy-regularized Bellman operator $\mathcal{T}_d^\pi$ in Eq. 7 and assume $\mathcal{H}(\mu^{\mathbf{s}_t,\mathbf{a}_t}, q_\theta^{\mathbf{s}_t,\mathbf{a}_t})$ is bounded for all $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$. We define $Q^{k+1} = \mathcal{T}_d^\pi Q^k$. Given q_θ, Q^{k+1} will converge to a corrected Q-value of π as $k \to \infty$ with the new objective function $J'(\pi)$ defined as*

$$J'(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\rho_\pi}\left[r(\mathbf{s}_t, \mathbf{a}_t) + \gamma f(\mathcal{H}(\mu^{\mathbf{s}_t,\mathbf{a}_t}, q_\theta^{\mathbf{s}_t,\mathbf{a}_t}))\right]. \tag{9}$$

The updating rule is $\pi_{\text{new}} = \arg\max_{\pi' \in \Pi} \mathbb{E}_{\mathbf{a}_t \sim \pi'}\left[Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t) + f(\mathcal{H}(\mu^{\mathbf{s}_t,\mathbf{a}_t}, q_\theta^{\mathbf{s}_t,\mathbf{a}_t}))\right]$ in phase of policy optimization. Next, we derive a new policy iteration algorithm, called *Distribution-Entropy-Regularized Policy Iteration (DERPI)*, alternating between policy evaluation and policy improvement. It provably converges to a policy regularized by the distribution-matching term.

Theorem 1. *(Distribution-Entropy-Regularized Policy Iteration) Repeatedly applying distribution-entropy-regularized policy evaluation in Eq. 7 and the policy improvement, the policy converges to an optimal policy π^* such that $Q^{\pi^*}(\mathbf{s}_t, \mathbf{a}_t) \geq Q^\pi(\mathbf{s}_t, \mathbf{a}_t)$ for all $\pi \in \Pi$.*

Please refer to Appendix F for the proof of Lemma 1 and Theorem 1. Theorem 1 demonstrates that if we incorporate the decomposed regularization into the actor critic in Eq. 9, we can design a variant of "soft policy iteration" [13] that can guarantee the convergence to an optimal policy given any fixed q_θ. In summary, our theoretical investigation is a variant of the standard analytical framework in MaxEnt RL that allows a comparable analysis. Importantly, we next recognize a fundamental difference between our decomposed entropy regularization and the vanilla entropy regularization in MaxEnt RL.

Uncertainty-aware Regularized Exploration in CDRL Compared with MaxEnt RL. For the objective function $J(\pi)$ in Eq. 6 of MaxEnt RL, the state-wise entropy $\mathcal{H}(\pi(\cdot|\mathbf{s}_t))$ is maximized explicitly *w.r.t.* π for policies with a higher entropy in terms of diverse actions to encourage an explicit exploration. For the objective function $J'(\pi)$ in Eq. 9 of distributional RL, the policy π is implicitly optimized through **the action selection process** $\mathbf{a}_t \sim \pi(\cdot|\mathbf{s}_t)$ guided by an augmented reward signal from the distribution-matching regularization $f(\mathcal{H}(\mu^{\mathbf{s}_t,\mathbf{a}_t}, q_\theta^{\mathbf{s}_t,\mathbf{a}_t}))$. Concretely, the learned policy is encouraged to visit state $\mathbf{s}_t$ along with the policy-determined action via $\mathbf{a}_t \sim \pi(\cdot|\mathbf{s}_t)$, whose current action-state return distributions $q_\theta^{\mathbf{s}_t,\mathbf{a}_t}$ *lag far behind* the target return distributions with

a large discrepancy. This discrepancy is measured by the magnitude of the cross entropy between two return distributions of $q_\theta^{\mathbf{s}_t,\mathbf{a}_t}$ and $\mu^{\mathbf{s}_t,\mathbf{a}_t}$. A large discrepancy indicates that the uncertainty of the current return distribution is considerably misestimated for the considered states, enabling an uncertainty-aware exploration against these states in the policy optimization phase. This also indicates that the policy learning in CDRL is additionally driven by the uncertainty difference between the current and the target estimates, leading to a distinct exploration strategy of distributional RL.

Interplay of Uncertainty-aware Regularization in Distributional Actor Critic. Putting the critic and actor learning together in distributional RL, we reveal their interplay impact pertinent to the uncertainty-aware regularized exploration. For the actor component, the policy learning seeks states and actions whose current return distribution estimate lags far behind the environmental uncertainty of the target returns. For the critic component, the critic learning reduces the return distribution mismatch on the states and actions explored by the policy, with two situations illustrated in Figure 2. This uncertainty-aware exploration effect arises from the decomposed regularization via the return density decomposition, interpreting the benefits of CDRL over classical RL.

5 Experiments

We comprehensively demonstrate our theoretical analysis using both Atari games and MuJoCo tasks. In this section, we validate that the uncertainty-aware regularization is crucial to the outperformance of CDRL over classical RL by varying ϵ in the return density decomposition. We also investigate the mutual impacts between the vanilla entropy regularization in MaxEnt RL and the uncertainty-aware entropy regularization from CDRL. Due to space limit, we provide the results in Appendix H.2. More implementation details, including the description of baselines, are provided in Appendix G.

Baseline Algorithm: $\mathcal{H}(\mu, q_\theta)(\varepsilon = 0.8/0.5/0.1)$. For the categorical distributional loss in C51 or the distributional critic loss in the actor critic, we employ $\widehat{\mu}^{s,a}$ instead of $\widehat{p}^{s,a}$ as the target return distribution, leading to the decomposed algorithms, denoted by $\mathcal{H}(\mu, q_\theta)$. This decomposed algorithm enables us to assess the uncertainty-aware regularization effect of distributional RL by directly comparing its performance with the classical RL and CDRL.

Experimental Details. We substantiate that the decomposed uncertainty-aware entropy regularization, derived in Eq. 4 through the return density function decomposition, plays a crucial role in the empirical superiority of CDRL over classical RL. We compare CDRL with the decomposed baseline algorithm $\mathcal{H}(\mu, q_\theta)$ under different ϵ based on Eq. 3. To ensure a pre-specified ϵ that guarantees a valid decomposition analyzed in Proposition 1, we employ a new notation ε, which is proportional to ϵ but is more convenient in the implementation. See Appendix G.1 for more explanation, including the transformation equation between ϵ and ε, and the details of the baseline algorithm $\mathcal{H}(\mu, q_\theta)$.

Results. Figure 3 showcases that as ε gradually decreases from 0.8 to 0.1, learning curves of decomposed C51, i.e., $\mathcal{H}(\mu, q_\theta)(\varepsilon = 0.8/0.5/0.1)$, tend to degrade from C51 to DQN across most Atari games. The sensitivity of the decomposed algorithm $\mathcal{H}(\mu, q_\theta)$ regarding ε depends on the environment. Similar results in MuJoCo environments can be found in Appendix H.1. Overall, our

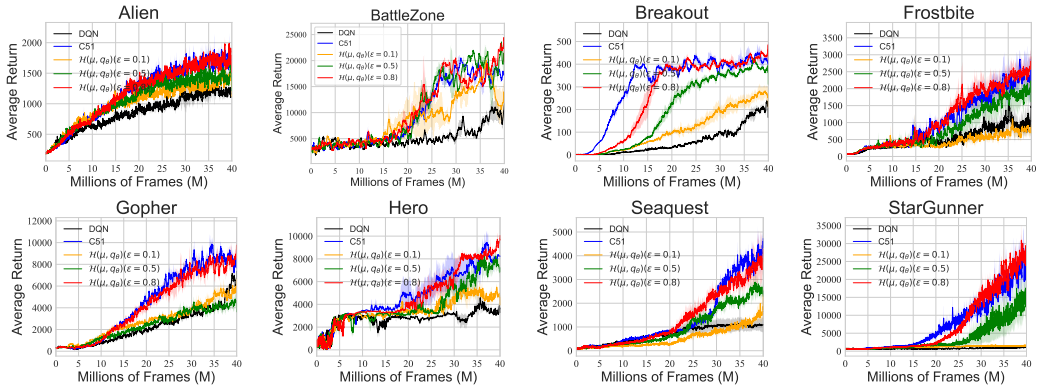

Figure 3: Learning curves of value-based CDRL (C51) and the decomposed algorithm $\mathcal{H}(\mu, q_\theta)(\varepsilon = 0.8/0.5/0.1)$ after applying the return distribution decomposition with different ε on eight Atari games. Results are averaged over three seeds, and the shade represents the standard deviation.

empirical result corroborates that the decomposed uncertainty-aware entropy regularization from the categorical distributional loss is pivotal to the empirical advantage of CDRL over classical RL.

6 Conclusion

In this study, we interpret the benefits of CDRL over classical RL as uncertainty-aware regularization via return density decomposition. In contrast to the exploration to encourage diverse actions in MaxEnt RL, the uncertainty-aware regularization in CDRL promotes exploring states where the environmental uncertainty is largely underestimated. Our study offers a novel exploration perspective to analyze the benefits of (categorical) distributional learning in RL. In future, it remains interesting yet challenging to extend our conclusion to general distributional RL, given that the analytical techniques, such as those in QR-DQN, are largely different from CDRL.

Acknowledgements

Bei Jiang and Linglong Kong were partially supported by grants from the Canada CIFAR AI Chairs program, the Alberta Machine Intelligence Institute (AMII), and Natural Sciences and Engineering Council of Canada (NSERC), and Linglong Kong was also partially supported by grants from the Canada Research Chair program from NSERC. We also thank all the constructive suggestions and comments from the reviewers.

References

[1] Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

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

[3] Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning*. MIT Press, 2023.

[4] 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*, 2017.

[5] Yu Chen, Xiangcheng Zhang, Siwei Wang, and Longbo Huang. Provable risk-sensitive distributional reinforcement learning with general function approximation. *International Conference on Machine Learning*, 2024.

[6] Taehyun Cho, Seungyub Han, Heesoo Lee, Kyungjae Lee, and Jungwoo Lee. Pitfall of optimism: Distributional reinforcement learning by randomizing risk criterion. *Advances in Neural Information Processing Systems*, 2023.

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

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

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

[10] Jesse Farebrother, Jordi Orbay, Quan Vuong, Adrien Ali Taïga, Yevgen Chebotar, Ted Xiao, Alex Irpan, Sergey Levine, Pablo Samuel Castro, Aleksandra Faust, et al. Stop regressing: Training value functions via classification for scalable deep rl. *arXiv preprint arXiv:2403.03950*, 2024.

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

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

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

[14] 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.

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

[16] Jianye Hao, Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Zhaopeng Meng, Peng Liu, and Zhen Wang. Exploration in deep reinforcement learning: From single-agent to multiagent domain. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NIPS Deep Learning Workshop*, 2015.

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

[19] Ehsan Imani and Martha White. Improving regression performance with distributional losses. In *International Conference on Machine Learning*, pages 2157–2166. PMLR, 2018.

[20] Haque Ishfaq, Guangyuan Wang, Sami Nur Islam, and Doina Precup. Langevin soft actor-critic: Efficient exploration through uncertainty-driven critic learning. *International Conference on Learning Representations*, 2015.

[21] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *International Conference on Learning Representations*, 2016.

[22] Qi Kuang, Zhoufan Zhu, Liwen Zhang, and Fan Zhou. Variance control for distributional reinforcement learning. *International Conference on Machine Learning*, 2023.

[23] Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, 2022.

[24] Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *International Conference on Machine Learning*, pages 6131–6141. PMLR, 2021.

[25] Shiau Hong Lim and Ilyas Malik. Distributional reinforcement learning for risk-sensitive policies. *Advances in Neural Information Processing Systems*, 35:30977–30989, 2022.

[26] Owen Lockwood and Mei Si. A review of uncertainty for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 155–162, 2022.

[27] Clare Lyle, Marc G Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4504–4511, 2019.

[28] 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.

[29] Yecheng Ma, Dinesh Jayaraman, and Osbert Bastani. Conservative offline distributional reinforcement learning. *Advances in neural information processing systems*, 34:19235–19247, 2021.

[30] 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.

[31] Alberto Maria Metelli, Amarildo Likmeta, and Marcello Restelli. Propagating uncertainty in reinforcement learning via wasserstein barycenters. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] 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.

[33] Mehryar Mohri. Foundations of machine learning, 2018.

[34] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *Neural Information Processing Systems (NeurIPS)*, 2019.

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

[36] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016.

[37] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016.

[38] Yangchen Pan, Kirby Banman, and Martha White. Fuzzy tiling activations: A simple approach to learning sparse representations online. *International Conference on Learning Representations*, 2019.

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

[40] 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*, pages 29–37. PMLR, 2018.

[41] 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.

[42] Mark Rowland, Rémi Munos, Mohammad Gheshlaghi Azar, Yunhao Tang, Georg Ostrovski, Anna Harutyunyan, Karl Tuyls, Marc G Bellemare, and Will Dabney. An analysis of quantile temporal-difference learning. *Journal of Machine Learning Research (JMLR)*, 2024.

[43] Mark Rowland, Yunhao Tang, Clare Lyle, Rémi Munos, Marc G Bellemare, and Will Dabney. The statistical benefits of quantile temporal-difference learning for value estimation. *International Conference on Machine Learning*, 2023.

[44] Mark Rowland, Li Kevin Wenliang, Rémi Munos, Clare Lyle, Yunhao Tang, and Will Dabney. Near-minimax-optimal distributional reinforcement learning with a generative model. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[45] Yang Sui, Yukun Huang, Hongtu Zhu, and Fan Zhou. Adversarial learning of distributional reinforcement learning. In *International Conference on Machine Learning*, pages 32783–32796. PMLR, 2023.

[46] Ke Sun, Bei Jiang, and Linglong Kong. How does return distribution in distributional reinforcement learning help optimization? In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2024.

[47] Ke Sun, Yi Liu, Yingnan Zhao, Hengshuai Yao, Shangling Jui, and Linglong Kong. Exploring the training robustness of distributional reinforcement learning against noisy state observations. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2023.

[48] Ke Sun, Yingnan Zhao, Yi Liu, Bei Jiang, and Linglong Kong. Distributional reinforcement learning with regularized wasserstein distance. *Advances in Neural Information Processing Systems*, 2024.

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

[50] Yunhao Tang and Shipra Agrawal. Exploration by distributional reinforcement learning. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[51] Kaiwen Wang, Owen Oertell, Alekh Agarwal, Nathan Kallus, and Wen Sun. More benefits of being distributional: Second-order bounds for reinforcement learning. *International Conference on Machine Learning*, 2024.

[52] Kaiwen Wang, Kevin Zhou, Runzhe Wu, Nathan Kallus, and Wen Sun. The benefits of being distributional: Small-loss bounds for reinforcement learning. *Advances in neural information processing systems*, 2023.

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

[54] Li Kevin Wenliang, Grégoire Déletang, Matthew Aitchison, Marcus Hutter, Anian Ruoss, Arthur Gretton, and Mark Rowland. Distributional bellman operators over mean embeddings. *International Conference on Machine Learning*, 2024.

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

[56] Harley Wiltzer, Marc G Bellemare, David Meger, Patrick Shafto, and Yash Jhaveri. Action gaps and advantages in continuous-time distributional reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[57] Harley Wiltzer, Jesse Farebrother, Arthur Gretton, and Mark Rowland. Foundations of multivariate distributional reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[58] Runzhe Wu, Masatoshi Uehara, and Wen Sun. Distributional offline policy evaluation with predictive error guarantees. *International Conference on Machine Learning*, 2023.

[59] 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.

[60] Liangyu Zhang, Yang Peng, Jiadong Liang, Wenhao Yang, and Zhihua Zhang. Estimation and inference in distributional reinforcement learning. *arXiv preprint arXiv:2309.17262*, 2023.

[61] Pushi Zhang, Xiaoyu Chen, Li Zhao, Wei Xiong, Tao Qin, and Tie-Yan Liu. Distributional reinforcement learning for multi-dimensional reward functions. *Advances in Neural Information Processing Systems*, 34:1519–1529, 2021.

A Related Work

Distributional Learning via Categorical Representation. Categorical learning has been widely employed, with advantages in representation [38, 21] and optimization [19, 46]. Recently, the empirical superiority of categorical distribution learning has been further investigated in various RL tasks [10]. A pressing need exists to examine the theoretical foundations of categorical distributional learning, particularly in RL. The perspective of uncertainty-aware regularized exploration that our study introduces provides significant insights into understanding the benefits of employing categorical distribution loss in the RL context.

Uncertainty-oriented Exploration. Uncertainty-oriented exploration plays an integral part in existing exploration methods [16], which leverages uncertainty either in the (posterior) estimation of the value function, as seen in Bayesian framework [37, 1], Bootstrap [36], and Ensemble methods [24], or in the entire distribution of returns [50, 30, 6]. For example, Decaying Left Truncated Variance (DLTV) [30] and Perturbed Quantile Regression (PQR) [6] exploit the variability of the learned return distribution to promote an optimistic exploration in distributional RL. In contrast, the primary aim of this study is to demonstrate that distributional learning in RL entails an intrinsic exploration effect against environmental uncertainty, contributing to the outperformance of distributional RL over classical RL. Our study goal is independent of designing advanced exploration strategies on top of distributional RL. Similarly, MaxEnt RL [55], which includes soft Q-learning [12], Soft Actor Critic (SAC) [13] and their variants [15], also promotes uncertainty-oriented exploration by relying on the stochasticity of the learned policy.

Uncertainty in RL. Uncertainty is ubiquitous in RL and sequential decision-making, and therefore harnessing uncertainty is always crucial in designing efficient algorithms [26]. In the literature of uncertainty quantification, uncertainty is often decomposed into two sources: *aleatoric uncertainty* and *epistemic uncertainty*.

- *Aleatoric uncertainty*, also called intrinsic or environmental uncertainty, originates from the stochastic or probabilistic nature of the environment, encompassing three main sources: stochastic transition dynamics, stochastic policy, and stochastic reward function. Aleatoric uncertainty is determined by the environment, which is thus irreducible. However, we can design more efficient algorithms by capturing more environmental uncertainty in the learning process, e.g., via distributional RL.
- *Epistemic uncertainty*, also called parametric uncertainty, often originates from the stochasticity in statistical estimation in the presence of limited data or incomplete knowledge. As opposed to aleatoric uncertainty, epistemic uncertainty is reducible and should decrease over more data, which contributes to a more reliable statistical estimation.

Uncertainty-oriented Exploration. There are a few survey papers that comprehensively summarize existing exploration approaches [23, 16]. Following [16], we classify the exploration strategies into two main categories: *uncertainty-oriented exploration* and *intrinsic motivation-oriented exploration*. The latter is inspired by psychology, which is not the focus of our study. Importantly, according to the two categories of uncertainty in RL, uncertainty-oriented exploration, which often applies *Optimism in the Face of Uncertainty* (OFU) principle, involves aleatoric and epistemic uncertainty.

- *Epistemic uncertainty-oriented exploration* takes advantage of the uncertainty in the (posterior) estimation of value functions. The typical exploration methods include Bayesian framework [37, 1, 31], Bootstrap [36], and Ensemble methods [24]. For instance, Bootstrapped DQN [36] maintains several independent Q-estimators and randomly samples one of them, enabling the agent to perform temporally extended exploration.
- *Aleatoric uncertainty-oriented exploration* aims to capture more environmental uncertainty from three sources of stochastic transition dynamics, stochastic policies, and stochastic reward function, all of which can be comprehensively integrated into return distribution. [6] employs Perturbed Quantile Regression (PQR) to promote the optimistic exploration within the distributional RL framework, while Decaying Left Truncated Variance (DLTV) [30] utilizes the variance from the learned return distributions. [50] investigates the approximate posterior sampling in distributional RL to encourage the exploration. By contrast, our primal goal in this study is to attribute the benefits of distributional RL to its intrinsic uncertainty-aware exploration we derived via return density decomposition instead of harnessing the

learned return distribution to develop subsequent aleatoric uncertainty-oriented exploration strategies in [50, 30]. On the other hand, MaxEnt RL [12, 13, 14] utilizes the stochasticity of learned policy, one of the three sources in environmental uncertainty, to encourage diverse actions. Therefore, MaxEnt RL can also be categorized into the aleatoric uncertainty-oriented exploration, and it is thus intuitive and interesting to make a detailed comparison of the exploration effects between distributional RL and MaxEnt RL, conducted in Section 4.1 of our study.

B More Details about Categorical Distributional RL and Algorithm Description of C51

Distributional Loss and Projection in CDRL. Categorical Distributional RL [2] uses the heuristic projection operator $\Pi_{\mathcal{C}}$, which was defined as

$$\Pi_{\mathcal{C}}\left(\delta_y\right)=\begin{cases}\delta_{z_1} & y \leq z_1 \\ \frac{z_{i+1}-y}{z_{i+1}-z_i}\delta_{l_i}+\frac{y-z_i}{z_{i+1}-z_i}\delta_{z_{i+1}} & z_i<y\leq z_{i+1} \\ \delta_{z_N} & y>z_N\end{cases}, \tag{10}$$

After applying the distributional Bellman operator $\mathfrak{T}^\pi$ on the current return distribution $\eta^\pi(s,a)$ in each update, the resulting new distribution, which we denote as $\widetilde{\eta}^\pi(s,a)$, typically no longer lies in the same (discrete) support with the original one on $\{z_i\}_{i=1}^N$. To maintain the same support, the underpinning of the KL divergence, CDRL additionally applies the projection operator $\Pi_{\mathcal{C}}$ on the new distribution $\widetilde{\eta}^\pi(s,a)$. This projection rule distributes the weight of δ_y across the original support points $\{z_i\}_{i=1}^N$ based on the linear interpolation. For example, if y lies in between two support points z_i and z_{i+1}, the probability mass on y is split between z_i and z_{i+1} with the weight inversely proportional to its distance ratio to z_i and z_{i+1}. Therefore, the projection extends affinely to finite mixtures of Dirac measures, such that for a mixture of Diracs $\sum_{i=1}^N p_i\delta_{y_i}$, we have $\Pi_{\mathcal{C}}\left(\sum_{i=1}^N p_i\delta_{y_i}\right)=\sum_{i=1}^N p_i\Pi_{\mathcal{C}}\left(\delta_{y_i}\right)$. The Cramér distance was recently studied as an alternative to the Wasserstein distances in the context of generative models [4]. Recall the definition of Cramér distance in the following.

Definition 1. *(Definition 3 [40]) The Cramér distance ℓ_2 between two distributions $\nu_1,\nu_2\in\mathscr{P}(\mathbb{R})$, with cumulative distribution functions F_{ν_1},F_{ν_2} respectively, is defined by:*

$$\ell_2\left(\nu_1,\nu_2\right)=\left(\int_{\mathbb{R}}\left(F_{\nu_1}(x)-F_{\nu_2}(x)\right)^2\,\mathrm{d}x\right)^{1/2}.$$

Further, the supremum-Cramér metric $\bar{\ell}_2$ is defined between two distribution functions $\eta,\mu\in\mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ by

$$\bar{\ell}_2(\eta,\mu)=\sup_{(x,a)\in\mathcal{X}\times\mathcal{A}}\ell_2\left(\eta^{(x,a)},\mu^{(x,a)}\right).$$

Thus, the contraction of categorical distributional RL can be guaranteed under Cramér distance:

Proposition 4. *(Proposition 2 [40]) The operator $\Pi_{\mathcal{C}}\mathcal{T}^\pi$ is a $\sqrt{\gamma}$-contraction in $\bar{\ell}_2$.*

An insight behind this conclusion is that Cramér distance endows a particular subset with a notion of orthogonal projection, and the orthogonal projection onto the subset is exactly the heuristic projection $\Pi_{\mathcal{C}}$ (Proposition 1 in [40]). [40] also states that the operator $\Pi_{\mathcal{C}}\mathcal{T}^\pi$ is contractive under Wasserstein distance.

Description of CDRL Algorithm: C51. With $N=51$, C51 instantiates the CDRL algorithm. To elaborate the algorithm, we first introduce the pushforward measure $f_{\#}\nu\in\mathcal{P}(\mathbb{R})$ from Definition 1 in [40]. This pushforward measure shifts the support of the probability measure μ according to the map f, which is commonly used in distributional RL literature. In particular, we consider an affine shift map $f_{r,\gamma}:\mathbb{R}\to\mathbb{R}$, defined by $f_{r,\gamma}(x)=r+\gamma x$. As Algorithm 1 displays, we first apply the pushforward measure on the target return distribution $\widehat{\eta}(s',a^*)$ by affinely shifting its support points, leading to a new distribution $\widetilde{\eta}(s,a)$. Next, we project the support points of $\widetilde{\eta}(s,a)$ by employing $\Pi_{\mathcal{C}}$ onto the original support, allowing us to compute the KL divergence in the end. Notably, we decompose the distributional objective function on the KL loss $\mathrm{KL}(\widehat{\eta}_{\text{target}}(s,a)||\widehat{\eta}(s,a))$.

Algorithm 1 CDRL Update (Adapted from Algorithm 1 in [40])

Require: Number of atoms N, e.g., $N = 51$ in C51, the categorical distribution $\widehat{\eta}(s,a) = \sum_{i=1}^{N} p_i^{s,a}\delta_{z_i}$ for the current return distribution.
Input: Sample transition (s, a, r, s')

```
if Policy evaluation: then
    a* ~ π(·|s')
else if Control: then
    a* ← arg max_{a'∈A} E_{R~η̂(s',a')}[R]
end if
η̃(s,a) ← (f_{r,γ})_# η̂(s',a*) # Distributional Bellmen update by applying T̂^π
η̂_target(s,a) ← Π_C η̃(s,a) # Project target support points and then distribute the probabilities
```

Output: Compute the distributional loss $\mathrm{KL}(\widehat{\eta}_{\text{target}}(s,a)||\widehat{\eta}(s,a))$ # Choose KL divergence as d_p

C Proof of Proposition 1

Proposition 1.(Decomposition Validity) Denote $\widehat{p}^{s,a}(x \in \Delta_E) = p_E/\Delta$, where p_E is the coefficient on the bin Δ_E. $\widehat{\mu}^{s,a}(x) = \sum_{i=1}^{N} p_i^{\mu}\mathbb{1}(x \in \Delta_i)/\Delta$ is a valid density if and only if $\epsilon \geq 1 - p_E$.

Proof. Recap a valid probability density function requires non-negative and one-bounded probability in each bin and all probabilities should sum to 1. We start to prove all probabilities should sum to 1, which is straightforward by taking the integral of both sides of Eq 3:

$$\begin{aligned}\int \widehat{p}^{s,a}(x)dx &= (1-\epsilon)\int \frac{\mathbb{1}\left(x \in \Delta_E\right)}{\Delta}dx + \epsilon\int \widehat{\mu}^{s,a}(x)dx \\ 1 &= (1-\epsilon) + \epsilon\int \widehat{\mu}^{s,a}(x)dx,\end{aligned} \tag{11}$$

which directly implies $\int \widehat{\mu}^{s,a}(x)dx = 1$. Next, we show necessity and sufficiency of non-negative and one-bounded probability in each bin.

Necessity. (1) When $x \in \Delta_E$, Eq. 3 can simplified as $p_E/\Delta = (1-\epsilon)/\Delta + \epsilon p_E^{\mu}/\Delta$, where $p_E^{\mu} = \widehat{\mu}(x \in \Delta_E)$. Thus, $p_E^{\mu} = \frac{p_E}{\epsilon} - \frac{1-\epsilon}{\epsilon} \geq 0$ if $\epsilon \geq 1 - p_E$. Obviously, $p_E^{\mu} = \frac{p_E}{\epsilon} - \frac{1-\epsilon}{\epsilon} \leq \frac{1}{\epsilon} - \frac{1-\epsilon}{\epsilon} = 1$ guaranteed by the validity of $\widehat{p}_E^{s,a}$. (2) When $x \notin \Delta_E$, we have $p_i/\Delta = \epsilon p_i^{\mu}/\Delta$, i.e.,When $x \notin \Delta_E$, We immediately have $p_i^{\mu} = \frac{p_i}{\epsilon} \leq \frac{1-p_E}{\epsilon} \leq 1$ when $\epsilon \geq 1 - p_E$. Also, $p_i^{\mu} = \frac{p_i}{\epsilon} \geq 0$.

Sufficiency. (1) When $x \in \Delta_E$, let $p_E^{\mu} = \frac{p_E}{\epsilon} - \frac{1-\epsilon}{\epsilon} \geq 0$, we have $\epsilon \geq 1 - p_E$. $p_E^{\mu} = \frac{p_E}{\epsilon} - \frac{1-\epsilon}{\epsilon} \leq 1$ in nature. (2) When $x \notin \Delta_E$, $p_i^{\mu} = \frac{p_i}{\epsilon} \geq 0$ in nature. Let $p_i^{\mu} = \frac{p_i}{\epsilon} \leq 1$, we have $p_i \leq \epsilon$. We need to take the intersection set of (1) and (2), and we find that $\epsilon \geq 1 - p_E \Rightarrow \epsilon \geq 1 - p_E \geq p_i$ that satisfies the condition in (2). Thus, the intersection set of (1) and (2) would be $\epsilon \geq 1 - p_E$.

In summary, as $\epsilon \geq 1 - p_E$ is both the necessary and sufficient condition, we have the conclusion that $\widehat{\mu}(x)$ is a valid probability density function $\iff \epsilon \geq 1 - p_E$.

□

D Proof of Proposition 2

Proposition 2 (Decomposed Neural FZI) Denote $q_\theta^{s,a}$ as the histogram density function of $Z_\theta^k(s,a)$ in Neural FZI. Based on Eq. 3 and KL divergence as d_p, Neural FZI in Eq. 2 is simplified as

$$Z_\theta^{k+1} = \underset{q_\theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}[\underbrace{-\log q_\theta^{s_i,a_i}(\Delta_E^i)}_{(a)} + \alpha\mathcal{H}(\widehat{\mu}^{s_i',\pi_Z(s_i')}, q_\theta^{s_i,a_i})]. \tag{12}$$

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

$$\begin{aligned} D_{\mathrm{KL}}(p, q_\theta) &= \sum_{i=1}^{N} \int_{l_{i-1}}^{l_i} \frac{p_i(x)}{\Delta} \log \frac{p^i(x)/\Delta}{q_\theta^i/\Delta} \, \mathrm{d}x \\ &= -\sum_{i=1}^{N} \int_{l_{i-1}}^{l_i} \frac{p_i(x)}{\Delta} \log \frac{q_\theta^i}{\Delta} \, \mathrm{d}x - \left(\sum_{i=1}^{N} \int_{l_{i-1}}^{l_i} \frac{p_i(x)}{\Delta} \log \frac{p^i(x)}{\Delta} \, \mathrm{d}x \right) \\ &= \mathcal{H}(p, q_\theta) - \mathcal{H}(p) \\ &\propto \mathcal{H}(p, q_\theta) \end{aligned} \tag{13}$$

where $p = \sum_{i=1}^{N} p_i(x) \mathbb{1}(x \in \Delta^i)/\Delta$ and $q_\theta = \sum_{i=1}^{N} q_i/\Delta$. 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 $\mathcal{R}(s_i, a_i) + \gamma Z_{\theta^*}^k(s_i', \pi_Z(s_i'))$. Then, we can derive the objective function within each Neural FZI as

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} \mathcal{H}(p^{s_i', \pi_Z(s_i')}, q_\theta^{s_i, a_i}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(-(1-\epsilon) \sum_{j=1}^{N} \int_{l_{j-1}}^{l_j} \frac{\mathbb{1}(x \in \Delta_E^i)}{\Delta} \log \frac{q_\theta^{s_i, a_i}(\Delta_j)}{\Delta} dx - \epsilon \sum_{j=1}^{N} \int_{l_{j-1}}^{l_j} \frac{p_j^\mu}{\Delta} \log \frac{q_\theta^{s_i, a_i}(\Delta_j)}{\Delta} dx \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left((1-\epsilon)(-\log q_\theta^{s_i, a_i}(\Delta_E^i)) + \epsilon \mathcal{H}(\widehat{\mu}^{s_i', \pi_Z(s_i')}, q_\theta^{s_i, a_i}) \right) + (1-\epsilon)\Delta \\ &\propto \frac{1}{n} \sum_{i=1}^{n} \left(-\log q_\theta^{s_i, a_i}(\Delta_E^i) + \alpha \mathcal{H}(\widehat{\mu}^{s_i', \pi_Z(s_i')}, q_\theta^{s_i, a_i}) \right), \text{ where } \alpha = \frac{\epsilon}{1-\epsilon} > 0 \end{aligned} \tag{14}$$

where recall that $\widehat{\mu}^{s_i', \pi_Z(s_i')} = \sum_{i=1}^{N} p_i^\mu(x) \mathbb{1}(x \in \Delta_i)/\Delta = \sum_{i=1}^{N} p_i^\mu/\Delta$ for conciseness and denote $q_\theta^{s_i, a_i} = \sum_{j=1}^{N} q_\theta^{s_i, a_i}(\Delta_j)/\Delta$. The cross-entropy $\mathcal{H}(\widehat{\mu}^{s_i', \pi_Z(s_i')}, q_\theta^{s_i, a_i})$ is based on the discrete distribution when $i = 1, ..., N$. Δ_E^i represent the interval that $\mathbb{E}\left[\mathcal{R}(s_i, a_i) + \gamma Z_{\theta^*}^k(s_i', \pi_Z(s_i'))\right]$ falls into, i.e., $\mathbb{E}\left[\mathcal{R}(s_i, a_i) + \gamma Z_{\theta^*}^k(s_i', \pi_Z(s_i'))\right] \in \Delta_E^i$. □

E Proof of Proposition 3

Proposition 3 (Equivalence between **the Mean-Related term** in Decomposed Neural FZI and Neural FQI) In Eq. 4, assume the function class $\{Z_\theta : \theta \in \Theta\}$ is sufficiently large such that it contains the target $\{Y_i^k\}_{i=1}^n$, when $\Delta \to 0$, for all k, minimizing **the mean-related term** in Eq. 4 implies

$$P(Z_\theta^{k+1}(s, a) = \mathcal{T}^{\mathrm{opt}} Q_{\theta^*}^k(s, a)) = 1, \quad \text{and} \quad \int_{-\infty}^{+\infty} \left| F_{q_\theta}(x) - F_{\delta_{\mathcal{T}^{\mathrm{opt}} Q_{\theta^*}^k(s,a)}}(x) \right| dx = o(\Delta), \tag{15}$$

where $\mathcal{T}^{\mathrm{opt}} Q_{\theta^*}^k(s, a)$ is the scalar-valued target in the k-th phase of Neural FQI, and $\delta_{\mathcal{T}^{\mathrm{opt}} Q_{\theta^*}^k(s,a)}$ is the Dirac delta function defined on the scalar $\mathcal{T}^{\mathrm{opt}} Q_{\theta^*}^k(s, a)$.

Proof. **Limiting Case.** Firstly, we define the distributional Bellman optimality operator $\mathfrak{T}^{\mathrm{opt}}$ as follows:

$$\mathfrak{T}^{\mathrm{opt}} Z(s, a) \stackrel{D}{=} \mathcal{R}(s, a) + \gamma Z\left(S', a^*\right), \tag{16}$$

where $S' \sim P(\cdot \mid s, a)$ and $a^* = \underset{a'}{\mathrm{argmax}} \mathbb{E}\left[Z\left(S', a'\right)\right]$. If $\{Z_\theta : \theta \in \Theta\}$ is sufficiently large enough such that it contains $\mathfrak{T}^{\mathrm{opt}} Z_{\theta^*}$ $(\{Y_i^k\}_{i=1}^n)$, then optimizing Neural FZI in Eq. 2 leads to $Z_\theta^{k+1} = \mathfrak{T}^{\mathrm{opt}} Z_{\theta^*}$.

Secondly, we apply the return density decomposition on the target histogram function $\widehat{p}^{s,a}(x)$. Consider the parameterized histogram density function h_θ and denote h_θ^E/Δ as the bin height in the

bin Δ_E, under the KL divergence between the first histogram function $\mathbb{1}(x \in \Delta_E)$ with $h_\theta(x)$, the objective function is simplified as

$$D_{\mathrm{KL}}(\mathbb{1}(x \in \Delta_E)/\Delta, h_\theta(x)) = -\int_{x \in \Delta_E} \frac{1}{\Delta} \log \frac{\frac{h_\theta^E}{\Delta}}{\frac{1}{\Delta}} dx \quad = -\log h_\theta^E \tag{17}$$

Since $\{Z_\theta : \theta \in \Theta\}$ is sufficiently large enough that can represent the pdf of $\{Y_i^k\}_{i=1}^n$, it also implies that $\{Z_\theta : \theta \in \Theta\}$ can represent the mean-related term part in its pdf via the return density decomposition. The KL minimizer would be $\widehat{h}_\theta = \mathbb{1}(x \in \Delta_E)/\Delta$ in expectation. Then, $\lim_{\Delta \to 0} \arg\min_{h_\theta} D_{\mathrm{KL}}(\mathbb{1}(x \in \Delta_E)/\Delta, h_\theta(x)) = \delta_{\mathbb{E}[Z^{\text{target}}(s,a)]}$, where $\delta_{\mathbb{E}[Z^{\text{target}}(s,a)]}$ is a Dirac Delta function centered at $\mathbb{E}\left[Z^{\text{target}}(s,a)\right]$ and can be viewed as a generalized probability density function. That being said, the limiting probability density function (pdf) converges to a Dirac delta function at $\mathbb{E}\left[Z^{\text{target}}(s,a)\right]$. In Neural FZI, we have $Z^{\text{target}} = \mathfrak{T}^{\text{opt}} Z_{\theta^*}$. Here, we use $Z_\theta^{k+1}(s,a)$ as the random variable whose cdf is the limiting distribution. According to the definition of the Dirac function, in the limiting case where $\Delta \to 0$, we attain that

$$\mathbb{P}(Z_\theta^{k+1}(s,a) = \mathbb{E}\left[\mathfrak{T}^{\text{opt}} Z_{\theta^*}^k(s,a)\right]) = 1. \tag{18}$$

This is because the pdf of the limiting return random variable $Z_\theta^{k+1}(s,a)$ is a Dirac delta function, which implies that the random variable takes this constant value with probability one. Due to the linearity of expectation in Lemma 4 of [2], we have

$$\mathbb{E}\left[\mathfrak{T}^{\text{opt}} Z_{\theta^*}^k(s,a)\right] = \mathfrak{T}^{\text{opt}} \mathbb{E}\left[Z_{\theta^*}^k(s,a)\right] = \mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a) \tag{19}$$

Finally, we obtain the convergence in probability one in the limiting case:

$$\mathbb{P}(Z_\theta^{k+1}(s,a) = \mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a)) = 1 \quad \text{as } \Delta \to 0 \tag{20}$$

Convergence in Distribution. The connection established above is in the limiting case. Alternatively, we can provide more formal proof by using the language of convergence in distribution. Here, we use $Z_{\theta,\Delta}^{k+1}$ to replace Z_θ^{k+1} to explicitly consider its asymptotic behavior. According to the fact that $\infty\{x \in \Delta_E\}/\Delta$ is the optimizer when minimizing the mean-related term in Eq. 4 given a fixed Δ, the convergence in distribution is:

$$\lim_{\Delta \to 0} \mathcal{D}(Z_{\theta,\Delta}^{k+1}) = \lim_{\Delta \to 0} \mathcal{D}(\mathbb{1}\{x \in \Delta_E\}/\Delta) = \mathcal{D}(\delta_{\mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a)}), \tag{21}$$

where $\delta_{\mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a)}$ is the Dirac Delta function centered at $\mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a)$. $\mathcal{D}(\delta_{\mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a)})$ is the corresponding step function, where $\mathcal{D}(\delta_{\mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a)})(x) = 1$ if $x \geq \mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a)$, and equals 0 otherwise. Note that the convergence in distribution in terms of the Dirac delta function implies that $\mathbb{P}(Z_\theta^{k+1}(s,a) = \mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a)) = 1$ as $\Delta \to 0$ in Eq 20.

Convergence Rate. In order to characterize how the difference varies when $\Delta \to 0$, we further define $\Delta_E = [l_e, l_{e+1})$ and we have:

$$\begin{aligned}
\int_{-\infty}^{+\infty} \left| F_{q_\theta}(x) - F_{\delta_{\mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a)}}(x) \right| dx &= \frac{1}{2\Delta} \left(\left(\mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a) - l_e \right)^2 + \left(l_{e+1} - \mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a) \right)^2 \right) \\
&= \frac{1}{2\Delta}(a^2 + (\Delta - a)^2) \\
&\leq \Delta/2 \\
&= o(\Delta),
\end{aligned} \tag{22}$$

where $\mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a) = \mathbb{E}\left[\mathfrak{T}^{\text{opt}} Z_{\theta^*}^k(s,a)\right] \in \Delta_E$ and we denote $a = \mathcal{T}^{\text{opt}} Q_{\theta^*}^k(s,a) - l_e$. The first equality holds as $q_\theta(x)$, the KL minimizer while minimizing the mean-related term, will follow a uniform distribution on Δ_E, i.e., $\widehat{q}_\theta = \mathbb{1}(x \in \Delta_E)/\Delta$. Thus, the integral of LHS would be the area of two centralized triangles accordingly. The inequality holds as the maximizer is obtained when $a = \Delta$ or 0. The result implies that the convergence rate in distribution difference is $o(\Delta)$.

□

F Convergence Proof of DERPI in Theorem 1

F.1 Proof of Distribution-Entropy-Regularized Policy Evaluation in Lemma 1

Lemma 1(Distribution-Entropy-Regularized Policy Evaluation) Consider the distribution-entropy-regularized Bellman operator $\mathcal{T}_d^\pi$ in Eq. 7 and assume $\mathcal{H}(\mu^{\mathbf{s}_t,\mathbf{a}_t}, q_\theta^{\mathbf{s}_t,\mathbf{a}_t})$ is bounded for all $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$. Define $Q^{k+1} = \mathcal{T}_d^\pi Q^k$, then Q^{k+1} will converge to a *corrected* Q-value of π as $k \to \infty$ with the new objective function $J'(\pi)$ defined as

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

Proof. Firstly, we plug in $V(\mathbf{s}_{t+1})$ into RHS of the iteration in Eq. 7, then we obtain

$$\begin{aligned}
&\mathcal{T}_d^\pi Q\left(\mathbf{s}_t, \mathbf{a}_t\right) \\
&= r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim P(\cdot|\mathbf{s}_t,\mathbf{a}_t)} \left[V\left(\mathbf{s}_{t+1}\right)\right] \\
&= r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim P(\cdot|\mathbf{s}_t,\mathbf{a}_t),\mathbf{a}_{t+1}\sim\pi} \left[f(\mathcal{H}(\mu^{\mathbf{s}_{t+1},\mathbf{a}_{t+1}}, q_\theta^{\mathbf{s}_{t+1},\mathbf{a}_{t+1}})) \right] + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim P(\cdot|\mathbf{s}_t,\mathbf{a}_t),\mathbf{a}_{t+1}\sim\pi} \left[Q\left(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}\right)\right] \\
&\triangleq r_\pi\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim P(\cdot|\mathbf{s}_t,\mathbf{a}_t),\mathbf{a}_{t+1}\sim\pi} \left[Q\left(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}\right)\right],
\end{aligned} \tag{23}$$

where $r_\pi\left(\mathbf{s}_t, \mathbf{a}_t\right) \triangleq r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim P(\cdot|\mathbf{s}_t,\mathbf{a}_t),\mathbf{a}_{t+1}\sim\pi} \left[f(\mathcal{H}(\mu^{\mathbf{s}_{t+1},\mathbf{a}_{t+1}}, q_\theta^{\mathbf{s}_{t+1},\mathbf{a}_{t+1}})) \right]$ is the entropy augmented reward. Applying the standard convergence results for policy evaluation [49], we can attain that this Bellman updating under $\mathcal{T}_d^\pi$ is convergent under the assumption of $|\mathcal{A}| < \infty$ and bounded entropy augmented rewards r_π. □

F.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 the policy optimization: $\pi_{new} = \arg\max_{\pi'\in\Pi} \mathbb{E}_{\mathbf{a}_t\sim\pi'} [Q^{\pi_{old}}(\mathbf{s}_t, \mathbf{a}_t) + f(\mathcal{H}\left(\mu^{\mathbf{s}_t,\mathbf{a}_t}, q_\theta^{\mathbf{s}_t,\mathbf{a}_t}\right))]$. Then $Q^{\pi_{new}}\left(\mathbf{s}_t, \mathbf{a}_t\right) \geq Q^{\pi_{old}}\left(\mathbf{s}_t, \mathbf{a}_t\right)$ for all $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$ with $|\mathcal{A}| < \infty$.*

Proof. The policy improvement in Lemma 2 implies that

$$\mathbb{E}_{\mathbf{a}_t\sim\pi_{\text{new}}} [Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t) + f(\mathcal{H}\left(\mu^{\mathbf{s}_t,\mathbf{a}_t}, q_\theta^{\mathbf{s}_t,\mathbf{a}_t}\right))] \geq \mathbb{E}_{\mathbf{a}_t\sim\pi_{\text{old}}} [Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t) + f(\mathcal{H}\left(\mu^{\mathbf{s}_t,\mathbf{a}_t}, q_\theta^{\mathbf{s}_t,\mathbf{a}_t}\right))].$$

We consider the Bellman equation via the distribution-entropy-regularized Bellman operator $\mathcal{T}_{sd}^\pi$:

$$\begin{aligned}
&Q^{\pi_{\text{old}}}\left(\mathbf{s}_t, \mathbf{a}_t\right) \\
&\triangleq r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim P} \left[V^{\pi_{\text{old}}}\left(\mathbf{s}_{t+1}\right)\right] \\
&= r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim P} \left[\mathbb{E}_{\mathbf{a}_{t+1}\sim\pi_{\text{old}}} \left[f(\mathcal{H}(\mu^{\mathbf{s}_{t+1},\mathbf{a}_{t+1}}, q_\theta^{\mathbf{s}_{t+1},\mathbf{a}_{t+1}})) + Q^{\pi_{\text{old}}}\left(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}\right) \right] \right] \\
&\leq r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim P} \left[\mathbb{E}_{\mathbf{a}_{t+1}\sim\pi_{\text{new}}} \left[f(\mathcal{H}(\mu^{\mathbf{s}_{t+1},\mathbf{a}_{t+1}}, q_\theta^{\mathbf{s}_{t+1},\mathbf{a}_{t+1}})) + Q^{\pi_{\text{old}}}\left(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}\right) \right] \right] \\
&= r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim P,\mathbf{a}_{t+1}\sim\pi_{\text{new}}} \left[f(\mathcal{H}(\mu^{\mathbf{s}_{t+1},\mathbf{a}_{t+1}}, q_\theta^{\mathbf{s}_{t+1},\mathbf{a}_{t+1}})) \right] + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim P,\mathbf{a}_{t+1}\sim\pi_{\text{new}}} [Q^{\pi_{\text{old}}}\left(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}\right)] \\
&= r_{\pi^{\text{new}}}\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim P,\mathbf{a}_{t+1}\sim\pi_{\text{new}}} [Q^{\pi_{\text{old}}}\left(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}\right)] \\
&\vdots \\
&\leq Q^{\pi_{\text{new}}}\left(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}\right),
\end{aligned} \tag{24}$$

where $Q^{\pi_{\text{old}}}\left(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}\right)$ indicates that the future actions are taking following π_{old}, given $\mathbf{s}_{s+t}$ and $\mathbf{a}_{t+1}$. We have repeated expanded $Q^{\pi_{\text{old}}}$ on the RHS by applying the distribution-entropy-regularized distributional Bellman operator. Each following step will then incorporate the actions following the new policy. Convergence to $Q^{\pi_{\text{new}}}$ follows from Lemma 1. □

F.3 Proof of DERPI in Theorem 1

Theorem 1 (Distribution-Entropy-Regularized Policy Iteration) Repeatedly applying distribution-entropy-regularized policy evaluation in Eq. 7 and the policy improvement, the policy converges to an optimal policy π^* such that $Q^{\pi^*}\left(\mathbf{s}_t, \mathbf{a}_t\right) \geq Q^\pi\left(\mathbf{s}_t, \mathbf{a}_t\right)$ for all $\pi \in \Pi$.

Proof. The proof is similar to soft policy iteration [13]. For completeness, we provide the proof here. By Lemma 2, as the number of iteration increases, the sequence Q^{π_i} at i-th iteration is monotonically increasing. Since we assume the uncertainty-aware entropy is bounded, 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:

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

According to the proof in Lemma 2, we can attain $Q^{\pi^*}(\mathbf{s}_t, \mathbf{a}_t) > Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$ for $(\mathbf{s}_t, \mathbf{a}_t)$. 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. □

G Implementation Details

G.1 Replacing ϵ with the ratio ε for Visualization

ε shares the same utility as ϵ, but it is more convenient in implementation. ε is defined as the mass proportion centered at the bin that contains the expectation *when transporting the mass to other bins*. A large proportion probability ε, which transports less mass to other bins, corresponds to a large ϵ in Eq. 3. Increasing ε indicates that the decomposed algorithm performs more similarly to a pure CDRL algorithm. As Proposition 1 elucidates, the return density decomposition requires that ϵ exceed certain thresholds to ensure the resultant decomposed $\widehat{\mu}^{s,a}$ qualifies as a valid density function. In practice, pinpointing this lower boundary for ϵ in each iteration to regulate its range could be prohibitively time-intensive. A more pragmatic approach involves redistributing the mass from the bin that contains the expectation to other bins in specified ratios, thereby introducing the corresponding ratio term ε. By varying ε from 0 to 1, it invariably meets the validity condition outlined in Proposition 1, thereby streamlining the process for conducting ablation studies concerning $\widehat{\mu}^{s,a}$ as demonstrated in Figure 3.

To delineate the relationship between the ratio ε and the coefficient ϵ in constructing $\widehat{\mu}^{s,a}$, after some calculations we establish their equivalence as follows:

$$\varepsilon = \frac{p_E - (1-\epsilon)}{p_E \epsilon}, \tag{25}$$

where p_E represents the weighting assigned to the bin Δ_E as specified in Proposition 1. The resulting $\varepsilon \in [0, 1]$ has a monotonically increasing relationship with ϵ. In addition, $\epsilon = 1$ implies $\varepsilon = 1$. These properties facilitate the visualization without undermining our conclusion.

Decomposition Details. By varying ε, we can evaluate ϵ via the transformation equation in Eq. 25, which guarantees the validity of return density decomposition. Next, under different ϵ, we compute the induced histogram density $\widehat{\mu}^{s,a}$ via the return density decomposition in Eq. 3:

$$\widehat{\mu}^{s,a}(x) = \widehat{p}^{s,a}(x \notin \Delta_E)/\epsilon + \widehat{p}^{s,a}(x \in \Delta_E)\varepsilon, \tag{26}$$

where combines Eq. 3 and Eq. 25. Importantly, by summing all the probabilities of p_i^{μ} in μ, we have:

$$\sum_{i=1}^{n} p_i^{\mu} = \frac{1 - p_E}{\epsilon} + \frac{p_E - (1-\epsilon)}{\epsilon} = 1. \tag{27}$$

This substantiates the validity of our decomposition by using ε instead of ϵ for visualization. Next, we replace $\widehat{p}^{s,a}$ with $\widehat{\mu}^{s,a}$ in C51 or the critic loss in Distributional AC (C51) as the decomposed algorithm $\mathcal{H}(\mu, q_\theta)$ and compare the performance of all considered algorithms. Please refer to the code in the implementation for more details.

G.2 Hyper-parameters and Network structure

Our implementation is adapted from the popular RLKit platform. For Distributional SAC with C51, we use 51 atoms similar to the C51 [2]. For distributional SAC with quantile regression, instead of using fixed quantiles in QR-DQN, we leverage the quantile fraction generation based on IQN [7] 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 ascending order. Besides, we adapt the sampling as $\tau_0 = 0, \tau_i = e_i / \sum_{i=0}^{N-1} e_i$, where $\epsilon_i \in U[0, 1], i = 1, ..., N$. We adopt the same hyper-parameters, which are listed in Table 1 and network structure as in the original distributional SAC paper [28].

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

H Experiments Results

H.1 Uncertainty-aware Regularization Effect by Varying ϵ in Actor Critic

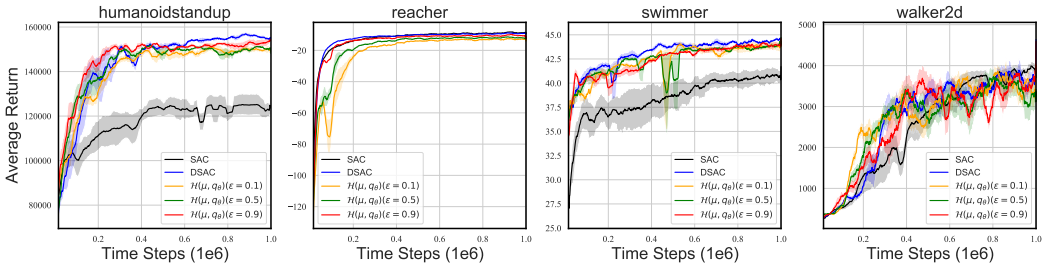

Figure 4: Learning curves of DSAC (C51) with the return distribution decomposition $\mathcal{H}(\mu, q_\theta)$ under different ε.

We study the uncertainty-aware regularization effect from being categorical distributional in the actor-critic framework, where we decompose the C51 critic loss in Distributional SAC (DSAC) according to Eq. 3. We denote the decomposed DSAC (C51) with different ε as $\mathcal{H}(\mu, q_\theta)(\varepsilon = 0.9/0.5/0.1)$.. As suggested in Figure 4, the performance of $\mathcal{H}(\mu, q_\theta)$ tends to vary from the vanilla DSAC (C51) to SAC with the decreasing of ε on four MuJoCo environments. In some environments, the difference of $\mathcal{H}(\mu, q_\theta)$ across various ε may not be pronounced between DSAC (C51) and SAC. We hypothesize that the algorithm performance is not sufficiently sensitive when ε changes within this restricted range. Although $\varepsilon \in (0, 1)$ is designed to guarantee a valid density decomposition, it does not guarantee that ϵ in Eq. 3 can flexibly vary from 0 to 1. It is worth noting that our return density decomposition is valid only when $\epsilon \geq 1 - p_E$ as shown in Proposition 1, and therefore ϵ can not strictly go to 0, where $\mathcal{H}(\mu, q_\theta)$ would degenerate to SAC ideally. Therefore, compared with the ablation study in Figure 3, the trend varying from DSAC to SAC in Figure 4 by decreasing ε may not be as pronounced as that in value-based RL evaluated on Atari games. One crucial reason behind is that the actor-critic architecture is generally perceived to be more prone to instability compared to value-based learning

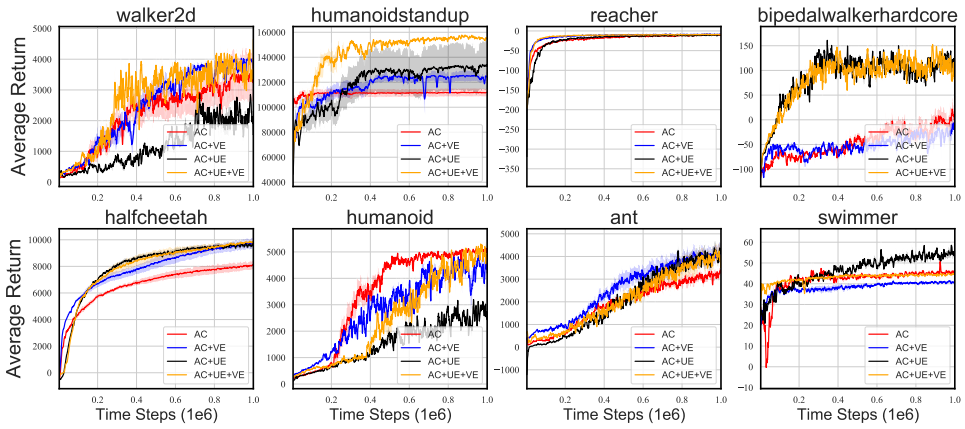

Figure 5: Learning curves of *AC*, *AC+VE* (SAC), ***AC+UE*** (DAC) and AC+UE+VE (DSAC) over five seeds across seven MuJoCo environments where the distributional RL part is based on C51. (**First Row**): Mutual Improvement. (**Second Row**): Potential Interference.

in RL. As outlined in [11], this instability stems from the policy updates, which likely introduces additional bias or variance from the critic learning process.

H.2 Mutual Impacts of the Two Entropy Regularization on DSAC (C51)

Baseline Algorithms. For a detailed comparison of the mutual impacts between **V**anilla **E**ntropy (**VE**) in MaxEnt RL and **U**ncertainty-aware **E**ntropy (**UE**) in CDRL, we conduct an ablation study across several related baseline algorithms. We denote SAC with/without vanilla entropy as *AC+VE* and *AC*. We denote Distributional SAC (DSAC) [28] with/without vanilla entropy as *AC+UE+VE* and ***AC+UE***. ***AC+UE*** is also denoted as *DAC*. The implementation details can be found in Appendix G.

Experimental Details. We demonstrate that the two types of regularized exploration in MaxEnt RL and CDRL play distinct roles in policy learning when employed simultaneously, including mutual improvement or potential interference. We perform our experiments for both DSAC (C51) in Figure 5 and DSAC (IQN) in Figure 6 of Appendix H.3, where the latter is used to examine the mutual impacts in quantile-based distributional RL heuristically.

Results. In the first row of Figure 5, simultaneously employing uncertainty-aware and vanilla entropy regularization renders a mutual improvement. Conversely, the two kinds of regularizations, when adopted together, can also lead to performance degradation, as exhibited in the second row in Figure 5. For instance, AC+UE+VE outperforms both *AC+VE* (SAC) and ***AC+UE*** (DAC) on humanoidstandup, while suffering from performance degradation on Ant and Swimmer. We posit that the potential interference may result from distinct exploration directions in the policy learning for the two types of regularizations. SAC optimizes the policy to visit states with high entropy, while distributional RL updates the policy to explore states and the associated actions whose current return distribution estimate lags far behind the environment uncertainty in target returns.

H.3 Mutual Impacts on DSAC (IQN)

To extend the mutual impact of the two types of regularization to broader distributional RL algorithm, we investigate the learning behavior of distributional RL based on IQN. The conclusion when using IQN is similar to that when using categorical distributional learning in Figure 5. In particular, in the first row of Figure 6, simultaneously employing uncertainty-aware and vanilla entropy regularization renders a mutual improvement. Conversely, the two kinds of regularizations, when adopted together, can also lead to performance degradation, as exhibited in the second row in Figure 6. For instance, on Swimmer and Reacher, *AC+UE+VE* is significantly inferior to ***AC+UE*** or *AC+VE*. These results about potential interference also serve as the emprical evidence to reveal distinct exploration directions in the policy learning for the two types of regularizations.

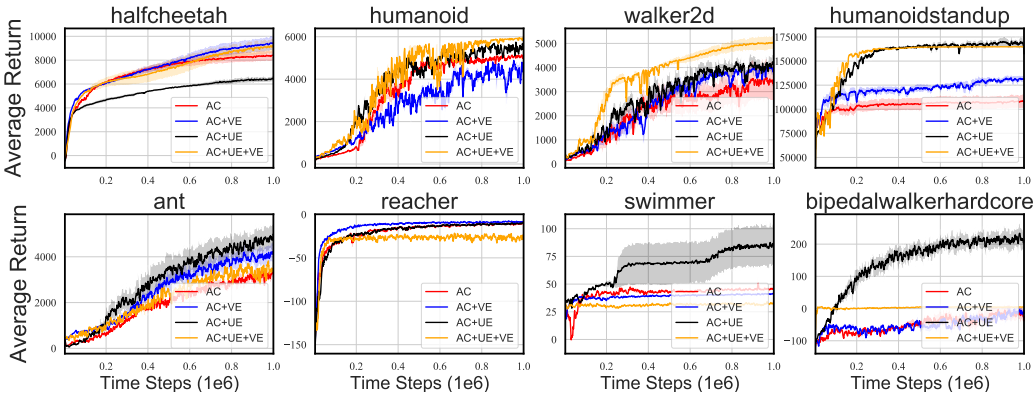

Figure 6: Learning curves of *AC*, *AC+VE* (SAC), ***AC+UE*** (DAC) and *AC+UE+VE* (DSAC) over five seeds across eight MuJoCo environments where DAC and DSAC are based on IQN. **(First Row)**: Mutual improvement. **(Second Row)**: Potential interference.

H.4 Ablation Study across Different Bin Sizes (Number of Atoms)

To further demonstrate our regularization effect based on the return density decomposition, we conducted an additional ablation study by varying the number of bins/atoms (equivalent to adjusting the bin sizes) of both C51 and our decompose algorithm $\mathcal{H}(\mu, q_\theta)$.

Figure 7 suggests that decreasing ε implies that $\mathcal{H}(\mu, q_\theta)$ degrades from C51 with the same bin size to DQN. Another interesting observation is that, as shown in Breakout (the first row in Figure 7), increasing the number of atoms (reducing the bin size) restricts the range of ϵ for a valid return density decomposition in Proposition 1. Consequently, a small number of atoms or a large bin size can allow a broader variation of $\mathcal{H}(\mu, q_\theta)$ from C51 to DQN, facilitating the demonstration of our regularization effect empirically.

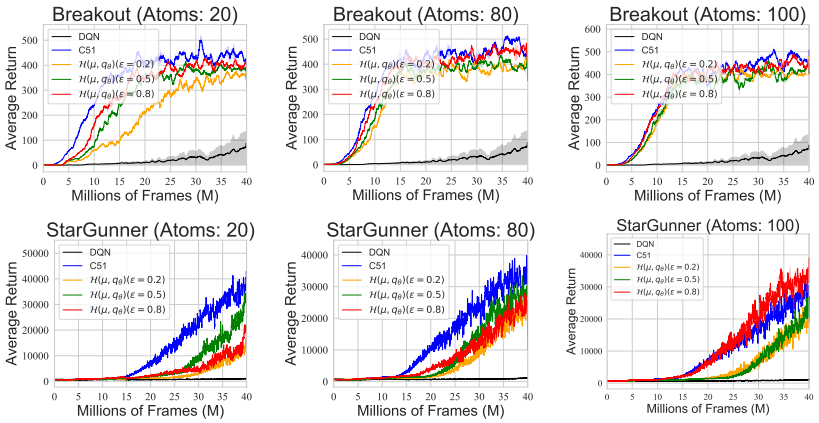

Figure 7: Learning curves of value-based CDRL, i.e., C51 algorithm, and the decomposed algorithm $\mathcal{H}(\mu, q_\theta)$ **across different numbers of atoms (various bin sizes)** on two Atari games. Results are averaged over three seeds, and the shade represents the standard deviation.