UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Benefits of (Categorical) Distributional Loss: Uncertainty-aware Regularized Exploration in Reinforcement Learning

Ke Sun, Yingnan Zhao, Enze Shi, Yafei Wang, Xiaodong Yan, Bei Jiang, Linglong Kong

University of Alberta
Alberta Machine Intelligence Institute (Amii)

NeurIPS 2025

Ke Sun

Outline

Reinforcement Learning is Increasingly Crucial

Games

Robotics

Transportation

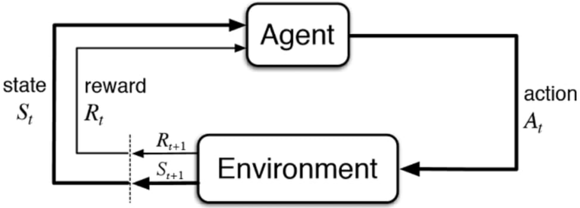
Healthcare

Economics

Language

Elements in Reinforcement Learning

- **Environment**: Markov Decision Process (MDP)



- **Return**: Cumulative Rewards (a random variable in nature)

$$Z^{\pi}(s,a)=\sum_{t=0}^{\infty}\gamma^{t}R(s_t,a_t), \tag{1}$$

where $a_t \sim \pi(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t,a_t), s_0 = s$, and $a_0 = a$.

Learning Principle: Reward Hypothesis

Reward Hypothesis

That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward).

Richard S. Sutton

Learning Expectation vs Distribution?

- Classical RL learns **value function**, the expectation of returns:

$$
\begin{aligned}
Q^{\pi}(s,a) &= \mathbb{E}\left[Z^{\pi}(s,a)\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s, a_0 = a\right]
\end{aligned}
$$

- Distributional RL learns the whole distribution of returns:

$$
\mathcal{D}(Z^{\pi}(s,a))
$$

where $\mathcal{D}$ extracts the distribution of a random variable.

Distributional Loss: Beyond Expectation

Fitted Q Iteration (FQI) vs Fitted Z Iteration (FZI)

- **Least Squares Loss in Classical RL**

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

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 and $Q_{\theta^*}^k$ is the target network updated between phases.

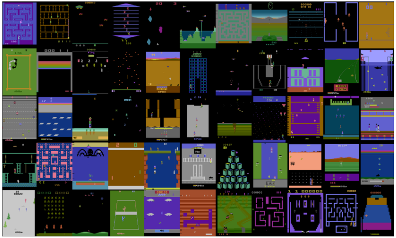
- **Distributional Loss in Distributional RL**

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

where $Y_i^k = \mathcal{R}(s_i, a_i) + \gamma Z_{\theta^*}^k\left(s_i', \pi_Z(s_i')\right)$ is the target return and π_Z follows the greedy rule $\pi_Z(s_i') = \operatorname{argmax}_{a'} \mathbb{E}\left[Z_{\theta^*}^k(s_i', a')\right]$. d_p is a distribution divergence / distance.

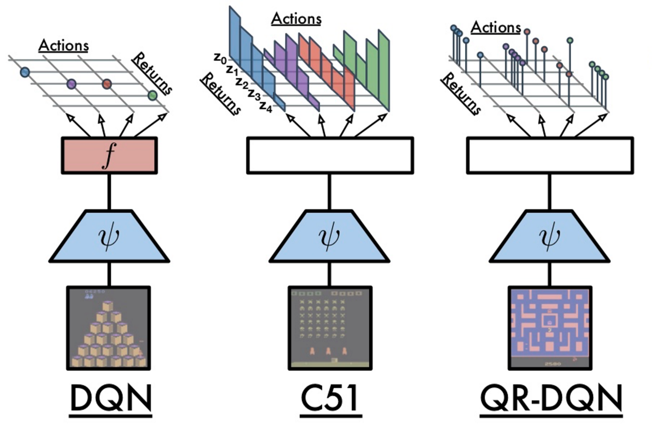
Performance Improvement of Distributional RL

Classical RL vs Distributional RL



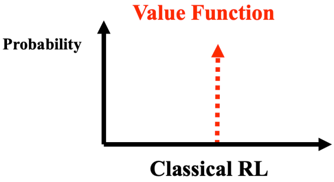
Atari Games

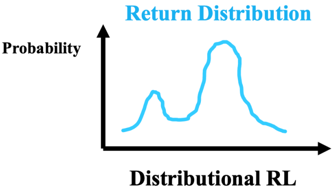
Existing Algorithms via Distribution Learning



A Fundamental Question

What are the **benefits** of distributional loss in RL?





Outline

Understand the Benefits of Distributional Loss

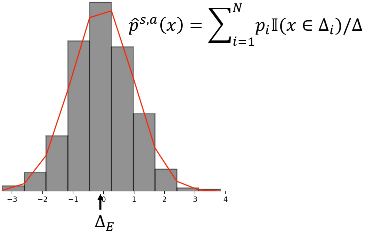
UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

1. **Key Technique**: Return Density Decomposition (inspired by gross error model in robust statistics)
2. **Value-based RL**:
   - Distribution-matching Entropy-regularized Loss Function
   - Asymptotic Connection with Least Squares Loss in Classical RL
   - Algorithm Difference: A New Entropy Regularization
3. **Policy-based RL**:
   - Connection with MaxEnt RL, e.g., Soft Actor Critic
   - Reward Augmentation and Uncertainty-aware Exploration

Outline

Histogram Density Estimator

- **Support Partition.** Given a fixed set of supports $l_0 \leq l_1 \leq \ldots \leq l_N$ with the equal bin size as Δ, each bin is thus denoted as $\Delta_i = [l_{i-1}, l_i), i = 1, \ldots, N-1$ with $\Delta_N = [l_{N-1}, l_N]$.
- **Histogram Density Estimator $\widehat{p}^{s,a}$.** $\widehat{p}^{s,a}$ with N bins is used to approximate an arbitrary continuous density $p^{s,a}$ of $Z^\pi(s,a)$: $\widehat{p}^{s,a}(x) = \sum_{i=1}^{N} p_i 1(x \in \Delta_i)/\Delta$. Δ_E as the interval that $\mathbb{E}\left[Z^\pi(s,a)\right]$ falls into, i.e., $\mathbb{E}\left[Z^\pi(s,a)\right] \in \Delta_E$.

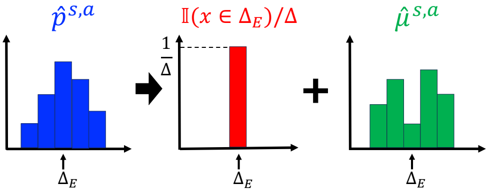


Return Density Decomposition

- **Return Density Decomposition.** We apply it on the histogram density function $\widehat{p}^{s,a}$ of the return $Z^\pi(s,a)$:

$$\widehat{p}^{s,a}(x) = (1-\epsilon)1(x \in \Delta_E)/\Delta + \epsilon\widehat{\mu}^{s,a}(x), \quad (4)$$

where **given** any $\widehat{p}^{s,a}$, $\widehat{\mu}^{s,a}$ is an **induced** histogram density function evaluated by $\widehat{\mu}^{s,a}(x) = \sum_{i=1}^{N} p_i^\mu 1(x \in \Delta_i)/\Delta$ with p_i^μ as the coefficient of the i-th bin Δ_i.



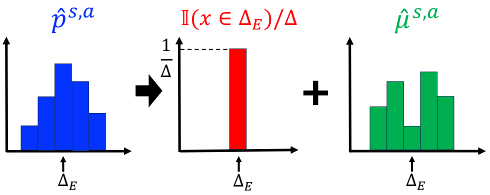
Decomposition Validity

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

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

Proposition 1. Decomposition Validity

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



Outline

Distributional RL: Entropy-regularized FQI

Fitted Q Iteration (FQI) vs Fitted Z Iteration (FZI)

- **Least Squares Loss in Classical RL**

$$Q_\theta^{k+1} = \text{argmin}_{Q_\theta} \frac{1}{n} \sum_{i=1}^{n} \left[y_i^k - Q_\theta(s_i, a_i) \right]^2, \tag{5}$$

where $y_i^k = r(s_i, a_i) + \gamma \max_{a \in \mathcal{A}} Q_{\theta^*}^k(s_i', a)$.

- **Distributional Loss in Distributional RL**

$$Z_\theta^{k+1} = \underset{Z_\theta}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} d_p(Y_i^k, Z_\theta(s_i, a_i)), \tag{6}$$

where $Y_i^k = \mathcal{R}(s_i, a_i) + \gamma Z_{\theta^*}^k(s_i', \pi_Z(s_i'))$ is the target return.

Next, we apply return density decomposition on Y_i^k and choose d_p as the KL divergence to rewrite the loss function.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Proposition 2. Decomposed Distributional Loss in FZI

Denote $q_\theta^{s,a}$ as the histogram density estimator of $Z_\theta^k(s,a)$ in FZI. Based on the return density decomposition and the KL divergence as d_p, the distributional loss in FZI is simplified as

$$Z_\theta^{k+1} = \underset{q_\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} [\underbrace{-\log q_\theta^{s_i,a_i}(\Delta_E^i)}_{\text{Mean-Related Term}} + \underbrace{\alpha \mathcal{H}(\widehat{\mu}^{s_i',\pi_Z(s_i')}, q_\theta^{s_i,a_i})}_{\text{Regularization Term}}], \quad (7)$$

where $\alpha = \varepsilon/(1-\varepsilon) > 0$ and the mean-related term is negative log-likelihood centered on Δ_E^i. $\mathcal{H}(p,q)$ is the cross-entropy between two probability density functions p and q.

Asymptotic Connection in Mean-Related Term

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Denote $\mathcal{T}^{\mathrm{opt}}$ as Bellman optimality operator $\mathcal{T}^{\mathrm{opt}} Q(s, a) = \mathbb{E}[\mathcal{R}(s, a)] + \gamma \max_{a'} \mathbb{E}_{s' \sim P}\left[Q\left(s', a'\right)\right]$.

Proposition 3. Equivalence between the Mean-Related term in Decomposed FZI and FQI

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 implies

$$\mathbb{P}(Z_\theta^{k+1}(s, a) = \mathcal{T}^{\mathrm{opt}} Q_{\theta^*}^k(s, a)) = 1, \tag{8}$$

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

Remark. Minimizing the mean-related term in the distributional loss in FZI is *asymptotically equivalent* to minimizing least squares loss in FQI with the same limiting minimizer.

Outline

Regularization Effect on Uncertainty

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- Environmental uncertainty represents the **whole stochasticity** in sequential decision-making:
  1. State transition.
  2. Reward function.
  3. Policy.
- In distributional RL, the histogram density function $\widehat{p}^{s_i,a_i}$ of Y_i^k captures the stochasticity of the **target return** (cumulative rewards over the trajectory) in each iteration.

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

- $\widehat{\mu}^{s,a}$ captures the uncertainty (higher-order moments information) of Y_i^k beyond the expectation.

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

Regularization Effect on Uncertainty

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- The Mean-Related term is asymptotically equivalent to learning the expectation in classical RL (by Proposition 3).

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

- Therefore, the Regularization term, which captures the higher-order moments information of the target return Y_i^k, is used to interpret the **benefits** of distributional loss over the least squares loss in classical RL.
- We call the regularization term as **uncertainty-aware regularization**, which is implicitly induced from distributional loss and we next show it promotes uncertainty-aware exploration in policy-based RL.

Equivalence to Categorical Distribution Loss

- **Categorical Distributional RL (CDRL)** is the first successful distributional RL family with the two components:
  1. Categorical distribution to represent the learned target return.
  2. d_p as the KL divergence.
- Histogram density function is equivalent to categorical distribution to represent a distribution given the aligned supports.
- Therefore, our analysis can be directly used to analyze the benefits of categorical distributional loss used in CDRL over classical RL.

Outline

Connection with MaxEnt RL

UNIVERSITY OF ALBERTA
EDMONTON · ALBERTA · CANADA

- **Explicit Regularization in MaxEnt RL.** MaxEnt RL *explicitly* encourages exploration by optimizing for policies (diverse actions) to reach states with higher entropy in the future:

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

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

- **Implicit Regularization in Distributional RL.** We apply return density decomposition in the (distributional) critic loss of actor-critic and focus on the regularization term. A new objective is

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

where as an extension, f can be any continuous increasing function over $\mathcal{H}$ and $\mu^{\mathbf{s}_t,\mathbf{a}_t}$ is derived after the decomposition.

Reward Augmentation in Actor Critic

- **Actor:** We optimize the policy π to maximize:

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

  where the augmented reward encourages the policy π 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 captured by $\mu^{\mathbf{s}_t, \mathbf{a}_t}$.
- **Critic:** The new objective is equivalent to a soft value function with a modified Bellman operator $\mathcal{T}_d^\pi$. Given a fixed q_θ, $\mathcal{T}_d^\pi$ is defined as

$$\mathcal{T}_d^\pi Q\left(\mathbf{s}_t, \mathbf{a}_t\right) \triangleq r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim P(\cdot|\mathbf{s}_t, \mathbf{a}_t)} \left[V\left(\mathbf{s}_{t+1}\right)\right], \quad (11)$$

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

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

Outline

Uncertainty-aware Regularized Exploration

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- **Exploration for Diverse Actions in MaxEnt RL.**

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

  where maximizing the shannon entropy simply encourages diversion actions to approach a uniform distribution.

- **Exploration for More Uncertain State in Distributional RL.**

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

  where the novel entropy derived from categorical distributional loss implicitly updates policies to **explore states with a large gap** between the true environmental uncertainty (approximated by $\mu^{\mathbf{s}_t, \mathbf{a}_t}$) and the current estimate $q_\theta^{\mathbf{s}_t, \mathbf{a}_t}$.

Interplay under Uncertainty-aware Regularization

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- **Actor:** The 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 (approximated by $\mu^{\mathbf{s}_t,\mathbf{a}_t}$) with a large discrepancy.
- **Critic:** $q_\theta^{s,a}$ is optimized to catch up with the uncertainty involved in the target return distribution of $\mu^{s,a}$, by minimizing the distributional loss d_p on all explored states and actions.

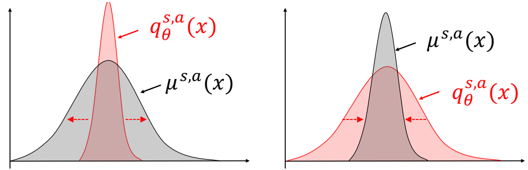


Figure: $q_\theta^{s,a}$ is optimized to disperse (left) or concentrate (right) to align with the uncertainty of target return distributions of $\mu^{s,a}$.

Outline

Experiments

We demonstrate two points:

① Regularization Effect of Distributional Loss on Performance

(sensitivity analysis by varying ϵ)

② Uncertainty-aware Regularization in Distributional RL vs Vanilla Entropy Regularization in MaxEnt RL

(ablation study)

Part 1: Regularization Effect on Performance

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- Recap the return density decomposition:

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

- **A Modified Algorithm:** $\mathcal{H}(\mu, q_\theta)(\varepsilon = 0.8/0.5/0.1)$.
  - We employ $\widehat{\mu}^{s,a}$ instead of $\widehat{p}^{s,a}$ as the target return distribution
  - We use $\mathcal{H}(\widehat{\mu}^{s,a}, q_\theta)$ instead of $d_p(\widehat{p}^{s,a}, q_\theta)$ to form the distributional loss.
  - 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.

Part 1: Regularization Effect by Varying ϵ

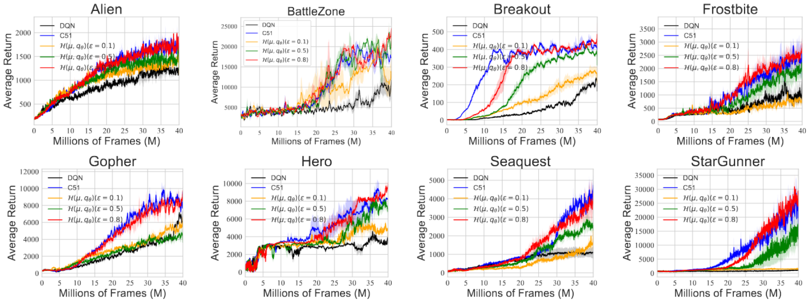


Figure: Learning curves of value-based CDRL (C51) and the decomposed algorithm $\mathcal{H}(\mu, q_\theta)(\varepsilon = 0.8/0.5/0.1)$ after applying the return distribution decomposition with different ε on eight Atari games.

Remark. $\mathcal{H}(\mu, q_\theta)$ interpolates between classical RL and distributional RL (CDRL). As ϵ decreases (less high-order moments distribution information), $\mathcal{H}(\mu, q_\theta)$ tends to the performance of classical RL.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Question: What is the **interplay** between uncertainty-aware regularization in distributional RL vs vanilla entropy regularization in MaxEnt RL?

- **Two Kinds of Regularization in Actor-Critic**
  1. **VE**: **V**anilla **E**ntropy regularization in MaxEnt RL or Soft Actor Critic (SAC)
  2. **UE**: **U**ncertainty-aware **E**ntropy regularization induced in categorical distributional loss in CDRL
- **Empirical Investigation via Ablation Study**
  1. *AC*: Actor Critic
  2. *AC+VE* : Actor Critic + vanilla entropy regularization ⇒ SAC
  3. ***AC+UE***: Distributional Actor Critic ⇒ DAC
  4. AC+UE+VE: Distributional Soft Actor Critic ⇒ DSAC

Part 2: Distributional RL vs MaxEnt RL

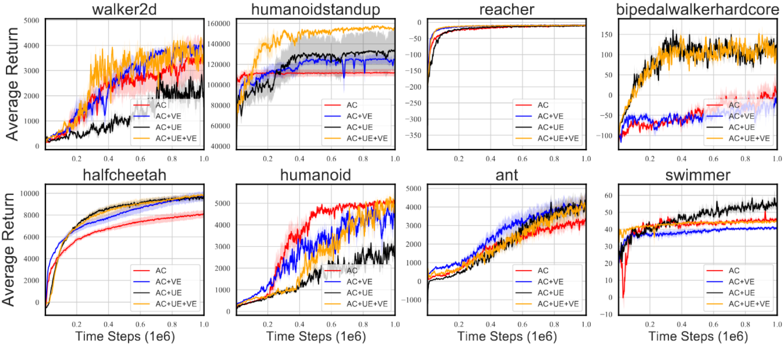


Figure: Learning curves of *AC*, *AC+VE* (SAC), ***AC+UE*** (DAC) and AC+UE+VE (DSAC) across eiggt MuJoCo environments where the distributional RL part is based on C51. (**First Row**): Mutual Improvement. (**Second Row**): Potential Interference.

Remark. The two regularizations have the effect of either mutual improvement or potential inference in distinct enviroments.

Outline

Extension to Quantile Distributional Loss

- d_p is often chosen as Wasserstein distance, which can be approximated by quantile regression in RL, such as Quantile Regression DQN, and Implicit Q Network (IQN).

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

- The quantile distributional loss can be viewed as a variant of composite quantile loss. It is also possible to decompose it into a mean-related term and a residual term.
- Minimizing the decomposed mean-related term is asymptotically mean-preserving as the number of quantiles approaches infinity inspired by quantile regression techniques (Some discussions are provided in Appendix M of our paper.)

Conclusion and Future Work

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Take-away Messages:

1. Try to use distributional loss instead of least squares loss in RL.
2. Distribution loss in RL learns more environmental uncertainty.
3. The benefit is an exploration bonus via an implicit regularization.

Open Problems and Future Work:

1. Benefits of distributional learning in RL with other distances, e.g., Wasserstein distance?
2. Other benefits of distributional learning in RL?
3. Distributional learning beyond RL, e.g., LLM, and the benefits?
4. When distributional learning may be harmful and why?

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Thank You!
Questions?