A Distributional Perspective on Reinforcement Learning

Presenter Qiaoling Li

School of Statistics and Management
Shanghai University of Finance and Economics, China

December 9, 2022

Outline

MOTIVATES I

- Classical value-based reinforcement learning methods attempt to model cumulative returns using expected values, expressed as state-value functions V(x) or action-value functions Q(x,a). In this modeling process, the complete distribution information is largely lost, and value distribution reinforcement learning is to solve this problem by modeling the distribution Z(x,a) of the random variable of cumulative return, rather than just modeling its expectations.

SETTING I

- $Q(x, a) = \mathbb{E}R(x, a) + \gamma\mathbb{E}Q(X', A')$
- $Z(x, a) \stackrel{D}{=} R(x, a) + \gamma Z(X', A')$

Bellman's Equations

- The return $Z^\pi = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t)$ is the sum of discounted rewards along the agent's trajectory of interactions with the environment.
- The value function Q^π of a policy π describes the expected return from taking action $a \in \mathcal{A}$ from state $x \in \mathcal{X}$.

$$Q^\pi(x, a) := \mathbb{E} Z^\pi(x, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(x_t, a_t)\right] \tag{1}$$

$$x_t \sim P(\cdot \mid x_{t-1}, a_{t-1}), a_t \sim \pi(\cdot \mid x_t), x_0 = x, a_0 = a \tag{2}$$

$$Q^\pi(x, a) = \mathbb{E} R(x, a) + \gamma \mathop{\mathbb{E}}_{P,\pi} Q^\pi(x', a'). \tag{3}$$

SETTING III

Bellman operator $\mathcal{T}^{\pi}$ and and optimality operator $\mathcal{T}$

$$\mathcal{T}^{\pi} Q(x, a) := \mathbb{E} R(x, a) + \gamma \underset{P, \pi}{\mathbb{E}} Q\left(x', a'\right) \tag{4}$$

$$\mathcal{T} Q(x, a) := \mathbb{E} R(x, a) + \gamma \mathbb{E}_{P} \max_{a' \in \mathcal{A}} Q\left(x', a'\right). \tag{5}$$

SETTING IV

The Wasserstein Metric

$$d_p(F,G) := \inf_{U,V} \|U - V\|_p \tag{6}$$

$$d_p(F,G) = \left\| F^{-1}(\mathcal{U}) - G^{-1}(\mathcal{U}) \right\|_p \tag{7}$$

$$d_p(F,G) = \left(\int_0^1 \left| F^{-1}(u) - G^{-1}(u) \right|^p du \right)^{1/p}. \tag{8}$$

SETTING V

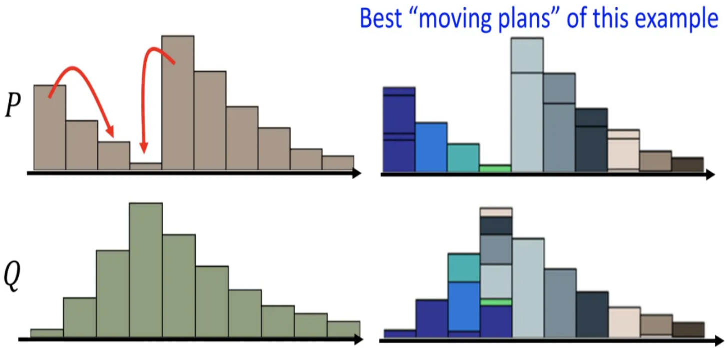


Figure 1: wasserstein-distance

SETTING VI

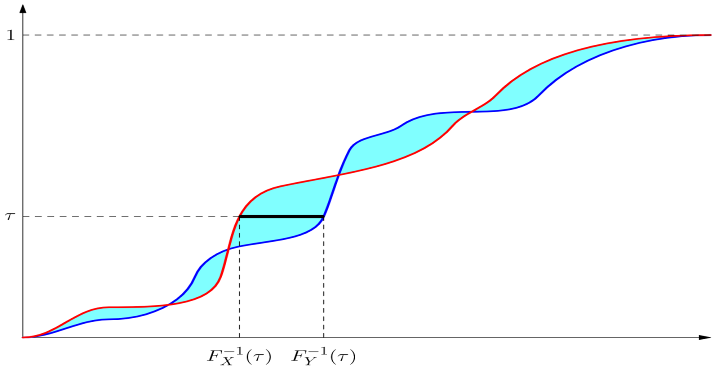


Figure 2: wasserstein-distance

SETTING VII

- where the infimum is taken over all pairs of random variables (U, V) with respective cumulative distributions F and G.
- The infimum is attained by the inverse c.d.f. transform of a random variable $\mathcal{U}$ uniformly distributed on $[0, 1]$

SETTING VIII

Properties

- Consider a scalar a and a random variable A independent of U, V . The metric d_p has the following properties:

$$d_p(aU, aV) \leq |a| d_p(U, V) \tag{9}$$

$$d_p(A + U, A + V) \leq d_p(U, V) \tag{10}$$

$$d_p(AU, AV) \leq \|A\|_p d_p(U, V). \tag{11}$$

- d_p is a metric over value distributions.
- For two value distributions $Z_1, Z_2 \in \mathcal{Z}$ we will make use of a maximal form of the Wasserstein metric:

$$\bar{d}_p\left(Z_1, Z_2\right) := \sup_{x,a} d_p\left(Z_1(x, a), Z_2(x, a)\right)$$

Policy Evaluation

- We view the reward function as a random vector $R \in \mathcal{Z}$, and define the transition operator $P^\pi : \mathcal{Z} \to \mathcal{Z}$

$$P^\pi Z(x,a) := Z\left(X', A'\right) \tag{12}$$

$$X' \sim P(\cdot \mid x,a), A' \sim \pi\left(\cdot \mid X'\right) \tag{13}$$

- We define the distributional Bellman operator $\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$as

$$\mathcal{T}^\pi Z(x,a) := R(x,a) + \gamma P^\pi Z(x,a) \tag{14}$$

Contraction

$$\mathcal{T}^{\pi} : \mathcal{Z} \to \mathcal{Z} \text{ is a } \gamma\text{-contraction in } \bar{d}_p \tag{15}$$

To prove this is to prove :

$$\bar{d}_p\left(\mathcal{T}^{\pi} Z_1, \mathcal{T}^{\pi} Z_2\right) \leq \gamma \bar{d}_p\left(Z_1, Z_2\right)$$

- Remark : π is fixed.

SETTING XI

Proof

Proof. Consider $Z_1, Z_2 \in \mathcal{Z}$. By definition,

$$\bar{d}_p\left(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2\right) = \sup_{x,a} d_p\left(\mathcal{T}^\pi Z_1(x,a), \mathcal{T}^\pi Z_2(x,a)\right)$$

By the properties of d_p, we have

$$\begin{aligned}
& d_p\left(\mathcal{T}^\pi Z_1(x,a), \mathcal{T}^\pi Z_2(x,a)\right) && (16) \\
& = d_p\left(R(x,a) + \gamma P^\pi Z_1(x,a), R(x,a) + \gamma P^\pi Z_2(x,a)\right) && (17) \\
& \leq \gamma d_p\left(P^\pi Z_1(x,a), P^\pi Z_2(x,a)\right) && (18) \\
& \leq \gamma \sup_{x',a'} d_p\left(Z_1\left(x',a'\right), Z_2\left(x',a'\right)\right) && (19)
\end{aligned}$$

SETTING XII

Control

A greedy policy π for $Z \in \mathcal{Z}$ maximizes the expectation of Z. The set of greedy policies for Z is:

$$\mathcal{G}_Z := \left\{ \pi : \sum_a \pi(a \mid x) \mathbb{E} Z(x, a) = \max_{a' \in \mathcal{A}} \mathbb{E} Z\left(x, a'\right) \right\} \tag{20}$$

We will call a distributional Bellman optimality operator any operator $\mathcal{T}$ which implements a greedy selection rule, i.e.

$$\mathcal{T} Z = \mathcal{T}^{\pi} Z$$

for some $\pi \in \mathcal{G}_Z$

SETTING XIII

lemma

Let $Z_1, Z_2 \in \mathcal{Z}$. Then

$$||\mathbb{E}\mathcal{T}Z_1 - \mathbb{E}\mathcal{T}Z_2||_\infty \leq \gamma||\mathbb{E}Z_1 - \mathbb{E}Z_2||_\infty$$

and in particular $\mathbb{E}Z_k \to Q^*$

SETTING XIV

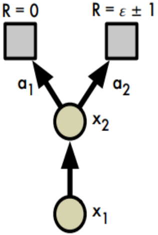
Proof.

Proof. The proof follows by linearity of expectation. Write $\mathcal{T}_D$ for the distributional operator and $\mathcal{T}_E$ for the usual operator. Then

$$\begin{aligned} \|\mathbb{E}\mathcal{T}_D Z_1 - \mathbb{E}\mathcal{T}_D Z_2\|_\infty &= \|\mathcal{T}_E \mathbb{E} Z_1 - \mathcal{T}_E \mathbb{E} Z_2\|_\infty && (21) \\ &\leq \gamma \|Z_1 - Z_2\|_\infty && (22) \end{aligned}$$

$\square$

SETTING XV



	x_1	x_2, a_1	x_2, a_2
Z^*	$\epsilon \pm 1$	0	$\epsilon \pm 1$
Z	$\epsilon \pm 1$	0	$-\epsilon \pm 1$
$\mathcal{T}Z$	0	0	$\epsilon \pm 1$

Figure 3: caption

Proof.

$$\bar{d}_1\left(Z, Z^*\right) = d_1\left(Z\left(x_2, a_2\right), Z^*\left(x_2, a_2\right)\right) = 2\epsilon$$

where we made use of the fact that $Z = Z^*$ everywhere except at (x_2, a_2). When we apply $\mathcal{T}$ to Z, however, the greedy action a_1 is selected and $\mathcal{T}Z\left(x_1\right) = Z\left(x_2, a_1\right)$. But

$$\bar{d}_1\left(\mathcal{T}Z, \mathcal{T}Z^*\right) = d_1\left(\mathcal{T}Z\left(x_1\right), Z^*\left(x_1\right)\right) \tag{23}$$

$$= \frac{1}{2}|1-\epsilon| + \frac{1}{2}|1+\epsilon| > 2\epsilon \tag{24}$$

for a sufficiently small ϵ. This shows that the undiscounted update is not a nonexpansion: $\bar{d}_1\left(\mathcal{T}Z, \mathcal{T}Z^*\right) > \bar{d}_1\left(Z, Z^*\right)$ □

C51 Algorithm I

C51 Algorithm

- We will model the value distribution using a discrete distribution parametrized by $N \in N$ and $V_{MIN}, V_{MAX} \in \mathbb{R}$, and whose support is the set of atoms

$$\{z_i = V_{MIN} + i\Delta Z : 0 \leq i < N\}$$

$\Delta z := \frac{V_{MAX} - V_{MIN}}{N-1}$

- The atom probabilities are given by a parametric model $\theta : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^N$

$$Z_\theta(x, a) = z_i$$

$\mathbf{w}.\mathbf{p}.p_i(x, a) := \frac{e^{\theta_i(x,a)}}{\sum_j e_j^\theta(x,a)}$

C51 Algorithm II

Project

$$\left(\Phi\hat{\mathcal{T}}Z_\theta(x,a)\right)_i = \sum_{j=0}^{N-1}\left[1-\frac{\left|\left[\hat{\mathcal{T}}z_j\right]_{V_{\mathrm{MIN}}}^{V_{\mathrm{Max}}}-z_i\right|}{\Delta z}\right]_0^1 p_j\left(x',\pi\left(x'\right)\right) \tag{25}$$

C51 Algorithm III

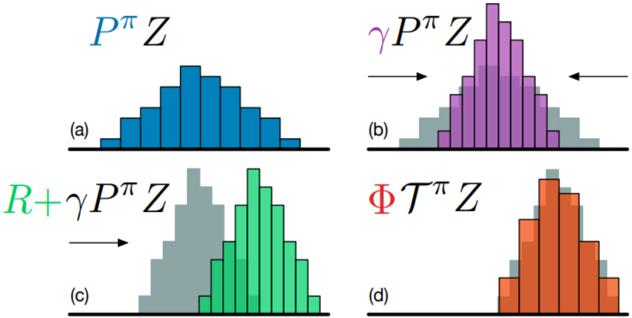


Figure 4: Bellman-operator

C51 Algorithm IV

Loss Function

- $D_{\mathrm{KL}}\left(\Phi \hat{\mathcal{T}} Z_{\tilde{\theta}}(x, a) \| Z_\theta(x, a)\right)$
- $\mathrm{KL}(p \| q)=\int p(x) \log \frac{p(x)}{q(x)} d x$
- $\mathrm{KL}(p \| q)=\sum_{i=1}^N p\left(x_i\right) \log \frac{p\left(x_i\right)}{q\left(x_i\right)}=\sum_{i=1}^N p\left(x_i\right)\left[\log p\left(x_i\right)-\log q\left(x_i\right)\right]$

C51 Algorithm V

Algorithm 1 Categorical Algorithm

input A transition $x_t, a_t, r_t, x_{t+1}, \gamma_t \in [0, 1]$

  $Q(x_{t+1}, a) := \sum_i z_i p_i(x_{t+1}, a)$

  $a^* \leftarrow \arg\max_a Q(x_{t+1}, a)$

  $m_i = 0, \quad i \in 0, \ldots, N-1$

  for $j \in 0, \ldots, N-1$ **do**

    # Compute the projection of $\hat{\mathcal{T}} z_j$ onto the support $\{z_i\}$

    $\hat{\mathcal{T}} z_j \leftarrow [r_t + \gamma_t z_j]_{V_{\text{MIN}}}^{V_{\text{MAX}}}$

    $b_j \leftarrow (\hat{\mathcal{T}} z_j - V_{\text{MIN}})/\Delta z$ # $b_j \in [0, N-1]$

    $l \leftarrow \lfloor b_j \rfloor, u \leftarrow \lceil b_j \rceil$

    # Distribute probability of $\hat{\mathcal{T}} z_j$

    $m_l \leftarrow m_l + p_j(x_{t+1}, a^*)(u - b_j)$

    $m_u \leftarrow m_u + p_j(x_{t+1}, a^*)(b_j - l)$

  end for

output $-\sum_i m_i \log p_i(x_t, a_t)$ # Cross-entropy loss

Figure 5: Caption

C51 Algorithm VI

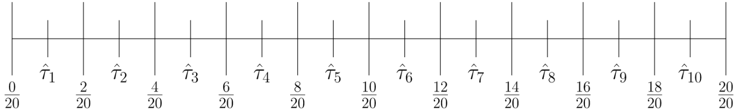
C51 VS DQN

- The framework of the C51 algorithm is still the DQN algorithm
- Use the ϵ-greedy policy
- The output of the convolutional neural network of the C51 algorithm is no longer a action-value function, but a probability at the fulcrum
- The loss function of the C51 algorithm is no longer the sum of mean squared deviations, but the KL divergence as described above

QR-DQN I

QR-DQN

The biggest difference between QR-DQN and C51 is the way the distribution is expressed.



$$N = 10 \qquad \hat{\tau}_i = \frac{2(i-1)+1}{2N} = \frac{2(i-1)+1}{20}$$

Figure 6: Quantile

QR-DQN II

Quantile-Regression

- $L_{MSE} = \min_{\beta} \sum_i^n (y_i - \mu(x_i, \beta))^2$
- $L_{MAE} = \sum_i^n |y_i - \xi(x_i, \beta)|$
- $L_{MAE} = \sum_{i:y_i \geq \xi(x_i,\beta)} (y_i - \xi(x_i, \beta_\tau)) + \sum_{i:y_i < \xi(x_i,\beta)} (\xi(x_i, \beta) - y_i)$
- $L_\tau = \sum_{i:y_i \geq \xi(x_i,\beta_\tau)} \tau (y_i - \xi(x_i, \beta_\tau)) + \sum_{i:y_i < \xi(x_i,\beta_\tau)} (1-\tau)(\xi(x_i, \beta_\tau) - y_i)$

QR-DQN III

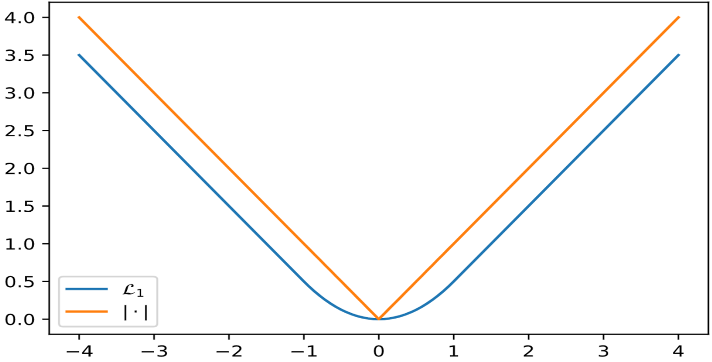


Figure 7: Quantile-Huber-Loss

QR-DQN IV

Loss function

$$L_\beta = \sum_{i=1}^{N} \mathbb{E}_Y \left[\rho^1_{\tau_i} \left(Y - \xi(\beta)_i \right) \right] \tag{26}$$

$$= \sum_{i=1}^{N} \mathbb{E}_{\mathcal{T}Z'} \left[\rho^1_{\hat{\tau}_i} \left(\mathcal{T}Z' - \theta_i \right) \right] \tag{27}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[\rho^1_{\hat{\tau}_i} \left(\mathcal{T}\theta'_j - \theta_i \right) \right] \tag{28}$$

QR-DQN V

Algorithm 1 Quantile Regression Q-Learning

Require: N, κ
input $x, a, r, x', \gamma \in [0, 1)$
  # Compute distributional Bellman target
  $Q(x', a') := \sum_j q_j \theta_j(x', a')$
  $a^* \leftarrow \arg\max_{a'} Q(x, a')$
  $\mathcal{T}\theta_j \leftarrow r + \gamma\theta_j(x', a^*), \quad \forall j$
  # Compute quantile regression loss (Equation 10)
output $\sum_{i=1}^{N} \mathbb{E}_j \left[\rho^{\kappa}_{\hat{\tau}_i}(\mathcal{T}\theta_j - \theta_i(x, a))\right]$

Figure 8: QR-DQN-Algorithm

QR-DQN VI

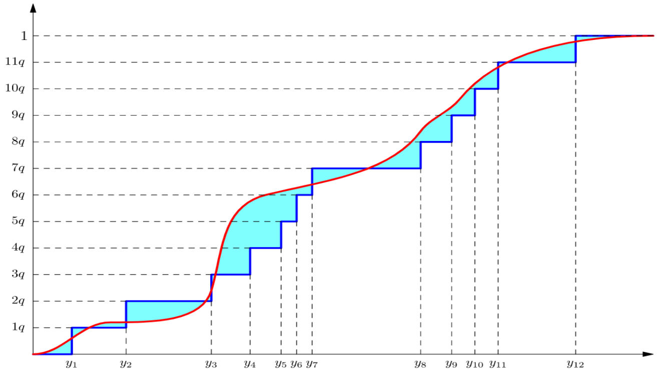


Figure 9: QR-DQN-1

QR-DQN VII

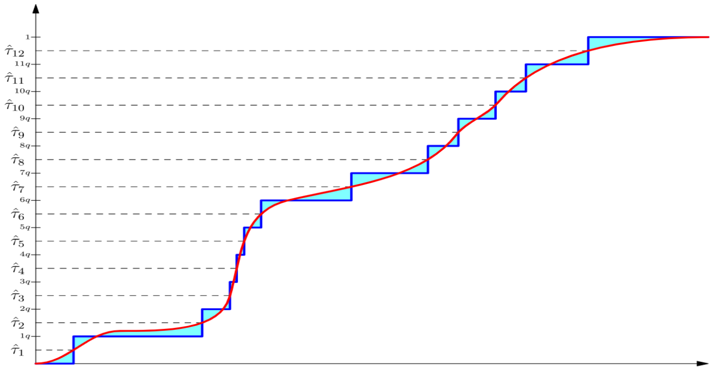


Figure 10: QR-DQN-2

QR-DQN VIII

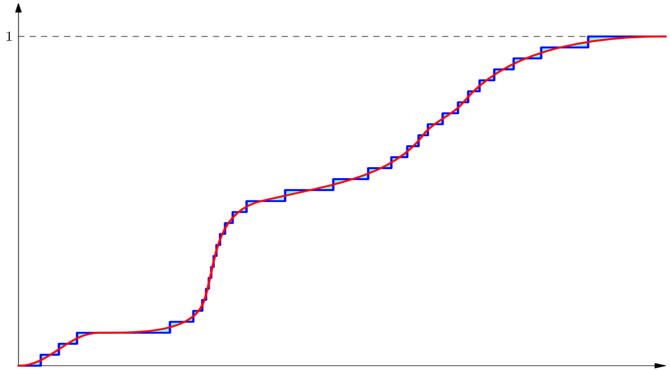


Figure 11: QR-DQN-3

QR-DQN IX

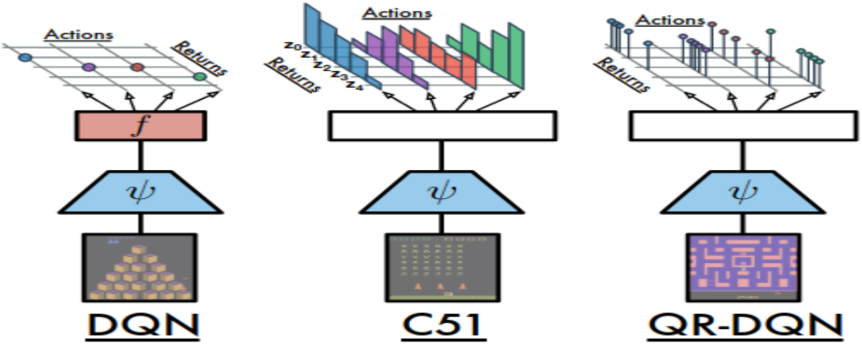


Figure 12: Caption

RESULTS I

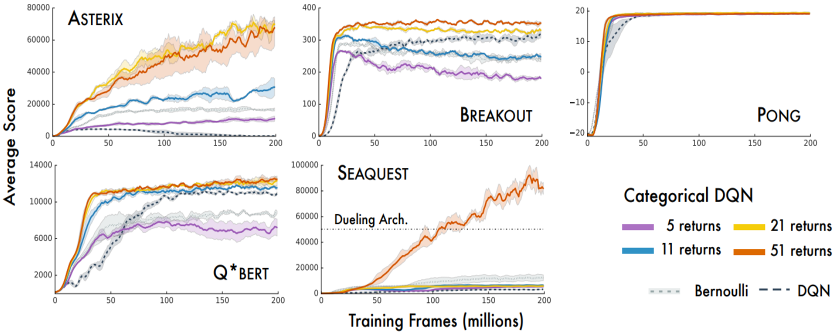


Figure 13: EXPERIMENTAL RESULTS

RESULTS II

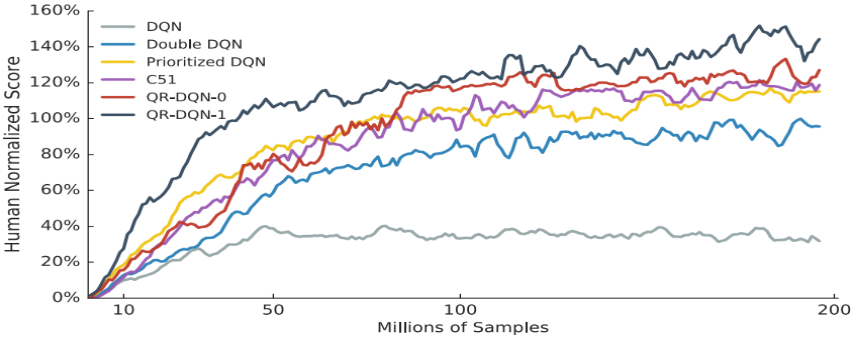


Figure 14: EXPERIMENTAL RESULTS

ACKNOWLEDGEMENT

Thank you all for your attention!