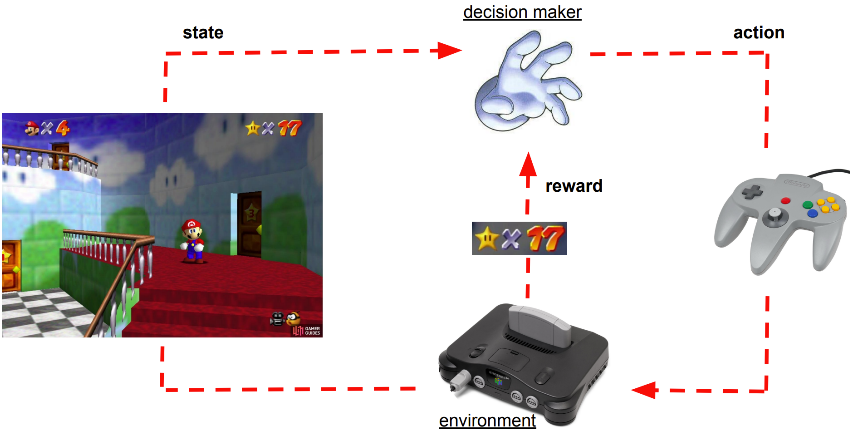
One-step distributional reinforcement learning

Mastane Achab

Technology Innovation Institute, Masdar City, Abu Dhabi, United Arab Emirates

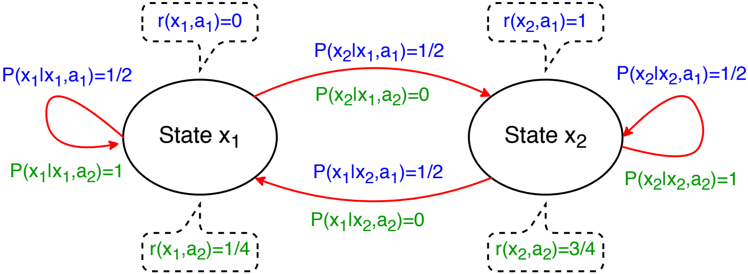
22 August 2023

Context: Sequential decision-making



Markov decision process (MDP)

An MDP [Puterman, 2014] is characterized by: states x, actions a, rewards $r(x, a, x')$ and transition probabilities $P(x'|x, a)$.



The control task

Optimality. Find a strategy π (mapping any state x to an action $\pi(x)$) that is optimal in terms of *expected* cumulative discounted return (for some discount factor $0 \leq \gamma < 1$):

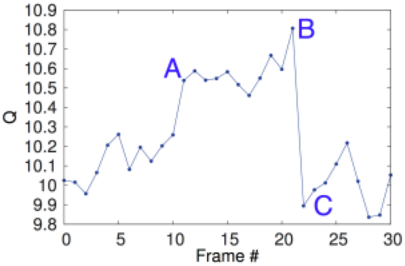
$$Q^*(x,a) = \max_{\pi} Q^{\pi}(x,a) := \mathbb{E}\left[\sum_{t\geq 0} \gamma^t r(X_t, A_t, X_{t+1}) \;\Big|\; X_0 = x, A_0 = a\right]$$

with states $X_{t+1} \sim P(\cdot|X_t, A_t)$ and actions $A_{t+1} = \pi(X_{t+1})$.

Reinforcement learning (RL). Learn an optimal strategy without knowing the transitions probabilities $P(x'|x,a)$ or the reward function: an RL agent only observes empirical transitions (x_t, a_t, r_t, x_{t+1}).

Deep Q-Network (DQN)

The DQN agent [Mnih et al., 2013] learns Q^* with a deep neural net Q_θ with parameters θ: successfully plays Atari games!

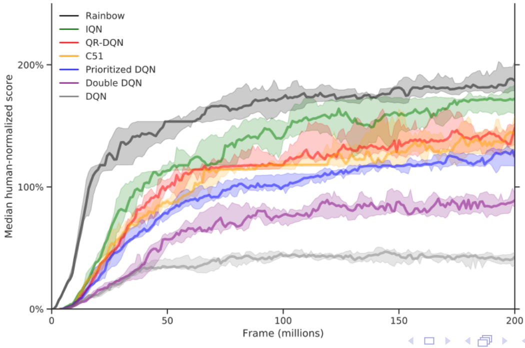


Distributional RL [Bellemare et al., 2017]

In distributional RL, the agent learns the whole probability distribution of the total return:

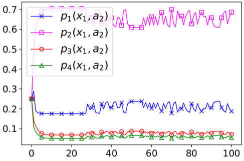
$$\mathrm{Law}\left(\sum_{t\geq 0} \gamma^t r(X_t, A_t, X_{t+1}) \;\middle|\; X_0 = x, A_0 = a; \pi\right).$$

In contrast, RL only focuses on the expected value $Q^\pi(x, a)$ of this distribution. On Atari games, distributional RL outperforms RL!

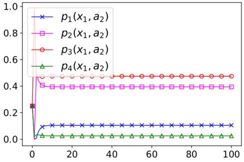


Our one-step solution to the instability of distributional RL

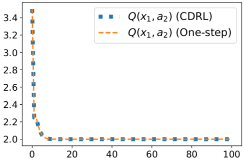
It has been shown that standard distributional RL algorithms are unstable for the control task [Bellemare et al., 2023].



(d) CDRL at (x_1, a_2).



(e) One-step at (x_1, a_2).



(f) Both Q-functions at (x_1, a_2).

$\rightarrow$ We solve this instability issue by only taking into account the randomness of the *one-step* dynamics!

	DistrRL	One-step DistrRL
Evaluation	$\mathrm{Distr}\left(\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t, X_{t+1})\right)$	$\sum_{x'} P(x'\|x,a)\delta_{r(x,a,x')+\gamma V^{\pi}(x')}$
Control	does not necessarily exist	$\sum_{x'} P(x'\|x,a)\delta_{r(x,a,x')+\gamma V^{*}(x')}$

Proposed tabular one-step categorical algorithm

We propose our one-step variant of tabular CDRL [Rowland et al., 2018].

Algorithm 1 Tabular one-step categorical DistrRL

Input: $\eta_t^{(x,a)} = \sum_{k=1}^{K} p_{t,k}(x,a)\delta_{z_k}$ for all (x,a)

Sample transition: (x_t, a_t, r_t, x_{t+1})

Estimate Q-values: $Q_t(x_{t+1}, a) \leftarrow \sum_{k=1}^{K} p_{t,k}(x_{t+1}, a) \cdot z_k$

if policy evaluation **then**

$\widehat{\eta}_t^{(x_t,a_t)} \leftarrow \Pi_{\mathcal{C}}(\delta_{r_t+\gamma \sum_{a'} \pi(a'|x_{t+1}) Q_t(x_{t+1},a')})$

else if control **then**

$\widehat{\eta}_t^{(x_t,a_t)} \leftarrow \Pi_{\mathcal{C}}(\delta_{r_t+\gamma \max_{a'} Q_t(x_{t+1},a')})$

end if

Mixture update: $\eta_{t+1}^{(x_t,a_t)} \leftarrow (1-\alpha_t(x_t,a_t))\eta_t^{(x_t,a_t)} + \alpha_t(x_t,a_t)\widehat{\eta}_t^{(x_t,a_t)}$

$\eta_{t+1}^{(x,a)} \leftarrow \eta_t^{(x,a)}$, $\forall (x,a) \neq (x_t, a_t)$

Output: η_{t+1}

Theorem (Convergence analysis [Achab et al., 2023])

Under standard Robbins-Monro condition, $\overline{W}_1(\eta_t, \eta_{\lim}) \xrightarrow{t\to\infty} 0$ *almost surely .*

Experiments - Atari video games

We test the one-step version of the C51 deep RL algorithm.

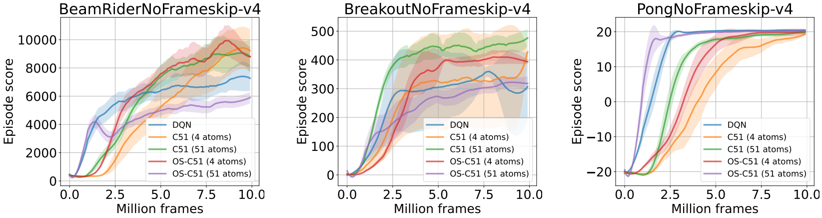
Algorithm 2 OS-C51 (single update)

Input: categorical distributions $\eta_\theta^{(x,a)} = \sum_{k=1}^{K} p_{\theta,k}(x,a)\delta_{z_k}$ and a transition (x_t, a_t, r_t, x_{t+1})

Compute Q-function in next state: $Q(x_{t+1}, a') \leftarrow \sum_{k=1}^{K} p_{\theta,k}(x_{t+1}, a')z_k$

Compute categorical target: $\widehat{\eta}^{(x_t,a_t)} \leftarrow \Pi_{\mathcal{C}}(\delta_{r_t+\gamma \max_{a'} Q(x_{t+1},a')})$

Output: $\mathrm{KL}(\widehat{\eta}^{(x_t,a_t)} \| \eta_\theta^{(x_t,a_t)})$



(a) Beamrider (b) Breakout (c) Pong

Figure: Performance on three Atari games.

References

Achab, M., ALAMI, R., DJILALI, Y. A. D., Fedyanin, K., and Moulines, E. (2023).
One-step distributional reinforcement learning.
Transactions on Machine Learning Research.

Bellemare, M. G., Dabney, W., and Munos, R. (2017).
A distributional perspective on reinforcement learning.
In *International Conference on Machine Learning*, pages 449–458. PMLR.

Bellemare, M. G., Dabney, W., and Rowland, M. (2023).
Distributional reinforcement learning.
MIT Press.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013).
Playing atari with deep reinforcement learning.
arXiv preprint arXiv:1312.5602.

Puterman, M. L. (2014).
Markov decision processes: discrete stochastic dynamic programming.
John Wiley & Sons.

Rowland, M., Bellemare, M., Dabney, W., Munos, R., and Teh, Y. W. (2018).
An analysis of categorical distributional reinforcement learning.
In *International Conference on Artificial Intelligence and Statistics*, pages 29–37. PMLR.