Distributional Reinforcement Learning for Efficient Exploration

Hengshuai Yao

Huawei Hi-Silicon

June, 2019

The exploration problem

- Exploration is a long standing problem in Reinforcement Learning.
- One major fundamental principle is optimism in the face of uncertainty.
- Both count-based methods and Bayesian methods follow this optimism principle.
- Here the uncertainty refers to *parametric uncertainty*, which arises from the variance in the estimates of certain parameters given finite samples.

Intrinsic uncertainties

- The estimation is not the only source of uncertainties.
- Most environment itself is stochastic.
- Even for deterministic game like GO, the opponent is a huge factor of uncertainty.
- The learning process can't eliminate Intrinsic uncertainty.

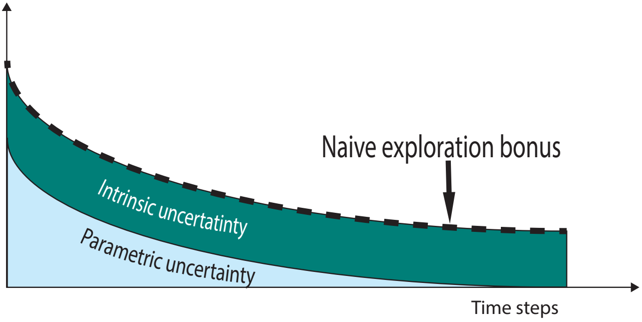
A naive approach

- A naive approach to exploration would be to use the variance of the estimated distribution as a bonus.
- Consider a multi-armed bandit environment with 10 arms where each arm's reward follows normal distribution $\mathcal{N}(\mu_k, \sigma_k)$.
- In the setting of multi-armed bandits, this approach leads to picking the arm a such that

$$a = \arg\max_k \bar{\mu}_k + c\sigma_k \tag{1}$$

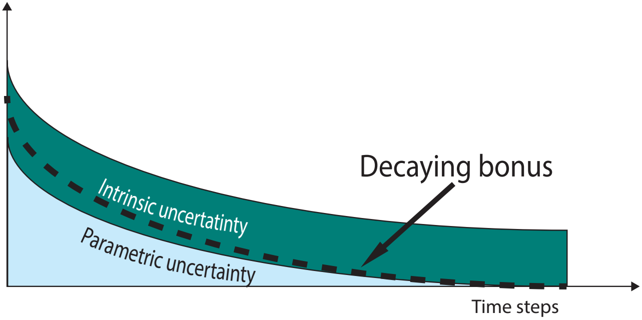
where $\bar{\mu}_k$ and σ_k^2 are the estimated mean and variance of the k-th arm, computed from the corresponding quantile distribution estimation.

The naive approach is not optimal



- The naive approach favors actions with high intrinsic uncertainty forever.

The motivation of Decaying exploration bonus



- To suppress the intrinsic uncertainty, we propose a decaying schedule in the form of a multiplier.

The DLTV exploration bonus

- For instantiating optimism in the face of uncertainty, the upper tail variability is more relevant than the lower tail.
- To increase stability, we use the left truncated measure of the variability, σ_+^2.
- By combining decaying schedule with σ_+^2 we obtain a new exploration bonus for picking an action, which we call Decaying Left Truncated Variance (DLTV):

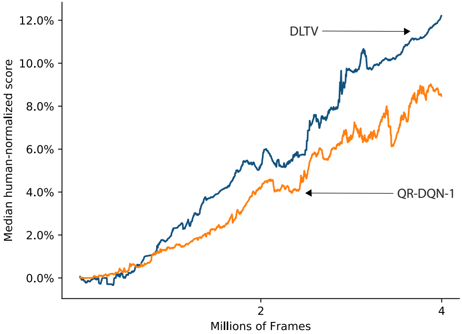
$$c_t\sqrt{\sigma_+^2}$$

where

$$c_t = c\sqrt{\frac{\log t}{t}}.$$

Results

- Our approach achieved 483 % average gain in cumulative rewards on the set of 49 Atari games.
- None of the learning curves exhibit plummeting behaviour
- Notably the performance gain is obtained in hard games such as Venture, PrivateEye, Montezuma Revenge and Seaquest.



Application on driving safety

- A particularly interesting application of the (Distributional) RL approach is driving safety.
- DLTV learns significantly faster than DQN and QR-DQN, achieving higher rewards for safety driving.

Summary

- Exploration is important.
- Principle is optimism in the face of uncertainty.
- Optimism without decaying is not optimal.
- Truncated measure is more stable.
- Combining them decaying schedule and truncated measure, we have DLTV.
- And it works.