Exploration in the face of Parametric and Intrinsic Uncertainties

Extended Abstract

Borislav Mavrin
Huawei Noah's Ark Lab, University of Alberta
mavrin@ualberta.ca

Shangtong Zhang
University of Oxford
zhangshangtong.cpp@gmail.com

Hengshuai Yao
Huawei Noah's Ark Lab
Edmonton, Alberta
hengshuai@gmail.com

Linglong Kong
Huawei Noah's Ark Lab, University of Alberta
lkong@ualberta.ca

ABSTRACT

In distributional reinforcement learning (RL), the estimated distribution of the value functions model both the parametric and intrinsic uncertainties. We propose a novel, efficient exploration method for Deep RL that has two components. The first is a decaying schedule to suppress the intrinsic uncertainty. The second is an exploration bonus calculated from the upper quantiles of the learned distribution. In Atari 2600 games, our method achieves 483 % average gain in cumulative rewards over QR-DQN.

KEYWORDS

distributional reinforcement learning; exploration

ACM Reference Format:

Borislav Mavrin, Shangtong Zhang, Hengshuai Yao, and Linglong Kong. 2019. Exploration in the face of Parametric and Intrinsic Uncertainties. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Exploration is a long standing problem in Reinforcement Learning (RL), where *optimism in the face of uncertainty* is one of the fundamental principles ([11, 16]). Here the uncertainty refers to *parametric uncertainty*, which arises from the variance in the estimates parameters due to finite samples. Both count-based methods ([1, 3, 9, 14, 18]) and Bayesian methods ([5, 9, 13]) follow this optimism principle. In this paper, we propose to use distributional RL methods to achieve this optimism.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada. © 2019 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

Different from classical RL methods, where an expectation of value function is learned ([12, 17, 19]), distributional RL methods ([4, 8]) maintain a full distribution of future return. In the limit, distributional RL captures the intrinsic uncertainty of an MDP ([4, 6, 7, 15]). *Intrinsic uncertainty arises from the stochasticity of the environment*, which is parameter and sample independent. However, during learning the estimated distribution is affected by both parametric and intrinsic uncertainties. It is not trivial how to separate these two. We propose an efficient approach of exploration that tries to isolate parametric uncertainty from the estimated distribution produced by Distributional RL.

2 BACKGROUND

As it was mentioned distributional RL focuses on learning the full distribution of the future return directly ([4, 8]). There are various approaches to represent a distribution in RL setting ([2, 4, 6]). In this paper, we focus on the quantile representation used in QR-DQN ([7]). In QR-DQN the distribution is learned via minimization of the following loss:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left[\rho^{\kappa}_{\hat{\tau}_i}\left(y_{t,i'}-\theta_i(s_t,a_t)\right)\right]$$

where θ_i is an estimation of the quantile corresponding to the quantile level (a.k.a. quantile index) $\hat{\tau}_i \doteq \frac{\tau_{i-1}+\tau_i}{2}$ with $\tau_i \doteq \frac{i}{N}$ for $0 \le i \le N$, $y_{t,i'} \doteq r_t + \gamma\theta_{i'}\left(s_{t+1}, \arg\max_{a'}\sum_{i=1}^{N}\theta_i(s_{t+1},a')\right)$, $\rho^{\kappa}_{\hat{\tau}_i}(x) \doteq |\hat{\tau}_i - \mathbb{I}\{x<0\}|\mathcal{L}_\kappa(x)$, $\mathcal{L}_\kappa$ is the Huber loss. θ_i can be parametrized by a neural network as in QR-DQN or by a single parameter as in multi-armed bandits. Therefore, the state-action value $Q(s,a)$ is simply the mean of $\{\theta_i\}_{i=1}^{N}$, i.e. $\frac{1}{N}\sum_{i=1}^{N}\theta_i(s,a)$. Similarly, the variance is $\frac{1}{N}\sum_{i=1}^{N}(\bar{\theta}-\theta_i)^2$.

3 ALGORITHM

A naive approach to exploration would be to use the variance of the estimated distribution as a bonus. We provide an illustrative counter

example. Consider a multi-armed bandit environment with 10 arms where each arm's reward follows normal distribution $\mathcal{N}(\mu_k, \sigma_k)$. In each run, means $\{\mu_k\}_k$ are drawn from standard normal. Standard deviation of the best arm is set to 1.0, other arms' standard deviations are set to 5. 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.

In this example naive exploration bonus fails. Specifically, the average reward is nearly zero after 3,000 steps averaged over 2,000 runs. The reason is that the estimated QR distribution is a mixture of parametric and intrinsic uncertainties. Recall, as learning progresses the parametric uncertainty vanishes and the intrinsic uncertainty stays. Therefore, this naive exploration bonus will tend to be biased towards the arm with high intrinsic variance but with low mean, which is not optimal.

The major obstacle in using the variance, i.e. σ_k^2 in (1) fr exploration, is the intractable interplay between parametric and intrinsic uncertainties in the estimated distribution. To suppress the intrinsic uncertainty, we propose a decaying schedule in the form of a multiplier to σ_k^2:

$$a = \arg\max_k \mu_k + c_t\sigma_k \tag{2}$$

From the classical QR theory ([10]), it is known that the parametric uncertainty of the quantile estimator decays at the following rate:

$$c_t = c\sqrt{\frac{\log t}{t}} \tag{3}$$

where c is a constant factor. This approach achieves average reward around 1 after 3,000 steps averaged over 2,000 runs.

We can improve the algorithm even further by making the following observation: QR has no restrictions on the family of distributions it can represent. In fact, the learned distribution can be *asymmetric*. The important question is how likely asymmetry can arise in applications. To test this hypothesis we measured the the difference between the mean and the median of the $\{\theta_i\}_{i=1}^N$ during training of QR-DQN in the game of Pong from Atari 2600 every 4,000 frames during 5M frames. The result is that the distribution is almost always asymmetric and asymmetry does not vanish as policy improves.

In order to account for asymmetry we propose to use the *Left Truncated Variance* (LTV) instead of the usual variance, i.e. σ^2. *Left Truncated Variance* is defined as:

$$\sigma_+^2 = \frac{2}{N}\sum_{i=\frac{N}{2}+1}^{N}(\bar{\theta} - \theta_i)^2 \tag{4}$$

Left truncation means that the left tail is truncated and we only consider the right tail. If the distribution is symmetric, then LTV is equal to the variance. However, in the case of asymmetric distribution they might not be equal and LTV would be biased towards the upper tail. In the multi-armed bandit testbed with asymmetric reward distributions LTV significantly outperforms the variance and has the same performance in the case of the symmetric reward distributions.

By combining the Decaying Schedule (3) and LTV (4) we propose a new exploration algorithm, Decaying Left Truncated Variance (DLTV):

$$a = \arg\max_k \mu_k + c_t\sigma_{+k} \tag{5}$$

DLTV generalizes in a straightforward fashion to Deep RL. Algorithm 1 outlines DLTV for Deep RL.

Algorithm 1 DLTV for Deep RL

Input: $w, w^-, (x, a, r, x'), \gamma \in [0, 1)$ ▷ network weights, sampled transition, discount factor

1: $Q(x', a') = \frac{1}{N}\sum_j \theta_j(x', a'; w^-)$
2: $a^* = \arg\max_{a'}(Q(x, a') + c_t\sqrt{\sigma_+^2})$
3: $\mathcal{T}\theta_j = r + \gamma\theta_j(x', a^*; w^-)$
4: $L(w) = \sum_i \frac{1}{N}\sum_j[\rho_{\hat{\tau}_i}(\mathcal{T}\theta_j - \theta_i(x, a; w))]$
5: $w' = \arg\min_w L(w)$

Output: w' ▷ Updated weights of $\theta()$

4 ATARI 2600 EXPERIMENTS

We evaluated DLTV on the set of 49 Atari games initially proposed by [12]. Algorithms were evaluated on 40 million frames, 3 runs per game. Our approach achieved 483 % average gain in cumulative rewards over QR-DQN. Notably the performance gain is obtained in hard games such as Venture, PrivateEye, Montezuma Revenge and Seaquest.

The architecture of the network follows [7]. For our experiments we chose the Huber loss with $\kappa = 1$ [1] as in the work by [7] due to its smoothness compared to $L1$ loss of QR-DQN-0. We followed closely [7] in setting the hyper parameters, except for the learning rate of the Adam optimizer which we set to $\alpha = 0.0001$.

The most significant distinction of our DLTV is the way the exploration is performed. *As opposed to QR-DQN there is no epsilon greedy exploration schedule in DLTV.* The exploration is performed via the σ_+^2 term only (line 2 of Algorithm 1).

An important hyper parameter which is introduced by DLTV is the schedule, i.e. the sequence of multipliers for σ_+^2, $\{c_t\}_t$. In our experiments we used the following schedule $c_t = 50\sqrt{\frac{\log t}{t}}$. We studied the effect of the decaying schedule in the Atari 2600 game Venture. Constant schedule with $c_t = 1, 5$ wasn't significantly different from the random agent. Whereas, DLTV achieves near human performance.

[1] QR-DQN with $\kappa = 1$ is denoted as QR-DQN-1

REFERENCES

[1] Peter Auer. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* (2002).

[2] Gabriel Barth-Maron, Matthew W Hoffman, David Budden, Will Dabney, Dan Horgan, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. 2018. Distributed distributional deterministic policy gradients. *arXiv preprint arXiv:1804.08617* (2018).

[3] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. 2016. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*.

[4] Marc G Bellemare, Will Dabney, and Rémi Munos. 2017. A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887* (2017).

[5] Richard Y Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. 2017. UCB Exploration via Q-Ensembles. *arXiv preprint arXiv:1706.01502* (2017).

[6] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. 2018. Implicit Quantile Networks for Distributional Reinforcement Learning. *arXiv preprint arXiv:1806.06923* (2018).

[7] Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. 2017. Distributional reinforcement learning with quantile regression. *arXiv preprint arXiv:1710.10044* (2017).

[8] Stratton C Jaquette. 1973. Markov decision processes with a new optimality criterion: Discrete time. *The Annals of Statistics* (1973).

[9] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. 2012. On Bayesian upper confidence bounds for bandit problems. In *Artificial Intelligence and Statistics*. 592–600.

[10] R. Koenker, A. Chesher, and M. Jackson. 2005. *Quantile Regression*. Cambridge University Press. https://books.google.ca/books?id=hdkt7V4NXsgC

[11] Tze Leung Lai and Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* (1985).

[12] 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. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529.

[13] Brendan O'Donoghue, Ian Osband, Remi Munos, and Volodymyr Mnih. 2017. The Uncertainty Bellman Equation and Exploration. *arXiv preprint arXiv:1709.05380* (2017).

[14] Georg Ostrovski, Marc G Bellemare, Aaron van den Oord, and Rémi Munos. 2017. Count-based exploration with neural density models. *arXiv preprint arXiv:1703.01310* (2017).

[15] Mark Rowland, Marc G Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. 2018. An Analysis of Categorical Distributional Reinforcement Learning. *arXiv preprint arXiv:1802.08163* (2018).

[16] Alexander L Strehl and Michael L Littman. 2005. A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd International Conference on Machine Learning*.

[17] Richard S Sutton. 1988. Learning to predict by the methods of temporal differences. *Machine Learning* (1988).

[18] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. 2017. # Exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems*.

[19] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine Learning* (1992).