Improving Sample-based Reinforcement Learning through Complex Non-parametric Distributions

Doctoral Consortium

Shi Yuan Tang
Nanyang Technological University
Singapore
shiyuan002@e.ntu.edu.sg

ABSTRACT

Sampling-based approaches in Reinforcement Learning (RL) typically involve learning or maintaining distributions. While many elegant algorithms were proposed in literature, most methods involve prior assumptions of the underlying distributions (eg. being Natural Exponential Family), or the number of modality for either simplicity or tractability reasons. A method to effectively apply complex or non-parametric distributions, for example, distributions approximated using neural network, is still lacking. One example is the limitation of using of reparameterized Gaussian policy, rather than any arbitrary non-parametric policy in Soft Actor-Critic (SAC) amenable to the necessary entropy estimation. The thesis would be focusing on proposing and evaluating methods to enable better approximation of complex distributions, and methods to estimate measurements of non-parametric distributions. The motivation is to allow better connections and applications of many deep learning and information theory techniques to the sampling-based approaches in RL, by alleviating limitations and difficulties of complex non-parametric distributions.

KEYWORDS

Sample-based Reinforcement Learning; Non-parametric Distributions; Distributional Reinforcement Learning; Risk-sensitive Reinforcement Learning; Soft Actor-Critic

ACM Reference Format:
Shi Yuan Tang. 2021. Improving Sample-based Reinforcement Learning through Complex Non-parametric Distributions: Doctoral Consortium. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 3 pages.

1 INTRODUCTION

Sampling can be applied in different parts of an RL algorithm. First, it can be applied in sampling from past experiences by maintaining a replay buffer. This application is commonly used in off-policy reinforcement learning (RL) through temporal difference learning to provide better sample efficiency and lower variance. Second, it can be used in population-based search or optimization algorithm of direct policy search methods, where a distribution is maintained for candidate solutions of the policy parameters. This involves learning the distribution based on an objective function. Monte Carlo approximation and importance sampling are proven to be viable strategies to approximate very complicated and analytically intractable functions. An extension to RL framework is through Markov Chain Monte Carlo (MCMC) or Cross-Entropy Method (CEM). Third, sampling could be used when a distribution is learned to model uncertainty. Some examples include modelling uncertainty during learning in Bayesian RL and modelling uncertainty in Q-value returned by the environment in distributional RL [1].

Particularly, the second and third applications mentioned involves maintaining and approximating distributions. For simplicity or tractability reasons, many related works involve prior assumptions of the underlying distributions. The representation capability and accuracy of the approximated distributions are thus inherently limited by parametric distributions. In addition, many information theory concepts or measurements such as entropy are theorized based on parametric distributions. Such statistical measurements are used for regularization in state-of-the-art RL algorithms to improve exploration and generalization. We propose the use of non-parametric distributions, replacing parametric distributions to address the restrictive representation capability and reduce imposed assumptions on the underlying distributions, with promising preliminary results. Additionally, a limitation in SAC [5] exists where policy with closed-form entropy like a uni-modal Gaussian is typically chosen. Future work could involve introducing an effective estimation method for the entropy of a non-parametric policy to enhance SAC's exploration behaviour in multi-modal problems.

2 PRELIMINARIES

The goal in RL is to learn a policy, which can be defined as the distribution over actions conditioned on states, $\pi(a_t \mid s_t)$. Another common way to define a policy is through Q-learning, where a state-action function is learned and implicitly used through an ϵ-greedy strategy. RL often uses a standard Markov Decision Process (MDP), modeled by the tuple $(\mathcal{S}, \mathcal{A}, R, T, \gamma)$, where $\mathcal{S}$ is a set of states $s \in \mathcal{S}$, $\mathcal{A}$ is a set of actions $a \in \mathcal{A}$, T defines transition distribution in the form $T(s_{t+1} \mid s_t, a_t)$ which describes the dynamics of the system, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defines the state- and action-dependent reward function, and $\gamma \in (0, 1]$ is a scalar discount factor. The aim of RL is to maximize the expectation of the sum of discounted returns:

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right] \mid a_t \sim \pi(\cdot|s_t), s_{t+1} \sim T(\cdot|s_t, a_t) \quad (1)$$

2.1 Direct Policy Search

For approaches where only action distribution is learned as the policy, policy gradient method (such as REINFORCE) or a direct

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

policy search (non-gradient based) method could be used (such as CEM). In CEM [7], the policy parameters $\boldsymbol{\phi}$ are generally sampled from a multivariate Gaussian distribution, $\boldsymbol{\phi} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma)$, where μ and Σ represent the mean and covariance. Each policy $\pi_{\boldsymbol{\phi}}$ is evaluated based on the rewards obtained from episodes sampled following the policy. After ranking the policies based on the episodic rewards, the elite policies $\pi_{\boldsymbol{\phi}[\text{elite}]}$ are identified by choosing the top ρ percentile of the sorted list of policies. The multivariate Gaussian policy distribution $\mathcal{N}(\mu, \Sigma)$ is then updated towards the distribution of the rare elite policies.

2.2 Q-learning & Actor-Critic

Instead of learning the distribution of actions, Q-learning aims to learn the state-action function, defined as $Q^{\pi}(s,a) = \mathbb{E}\left[Z^{\pi}(s,a)\right]$, where random variable $Z^{\pi}(s,a) = \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)$ is the sum of discounted returns. When θ represents the parameters of Q-function, the estimate could be improved by repeatedly applying the Bellman update:

$$Q_\theta(s_t, a_t) \leftarrow \mathbb{E}[R(s_t, a_t)] + \gamma \mathbb{E}_T \left[\max_{a_{t+1}} Q_\theta\left(s_{t+1}, a_{t+1}\right) \right] \quad (2)$$

The Q-function is approximated by minimizing the temporal difference error, and θ is updated accordingly,

$$\delta_t = r_t + \gamma \max_{a_{t+1} \in \mathcal{A}} Q_\theta\left(s_{t+1}, a_{t+1}\right) - Q_\theta\left(s_t, a_t\right) \quad (3)$$

$$\theta \leftarrow \theta + \alpha_\theta \delta_t \nabla_\theta Q_\theta(s_t, a_t) \quad (4)$$

In actor-critic framework, a separate policy (or actor) is learned in conjunction with the Q-function (critic). The policy is updated in the direction suggested by the critic using policy gradients,

$$\phi \leftarrow \phi + \alpha_\phi Q_\theta(s_t, a_t) \nabla_\phi \log \pi_\phi(a_t \mid s_t) \quad (5)$$

3 DIFFERENT ROLES OF DISTRIBUTION

The thesis will explore applying non-parametric distributions to facilitate search, exploration, uncertainty and improve learning generalisation in RL. We could categorize the incorporation of learning distributions in the following RL components:

3.1 Distributions of Policy Parameters

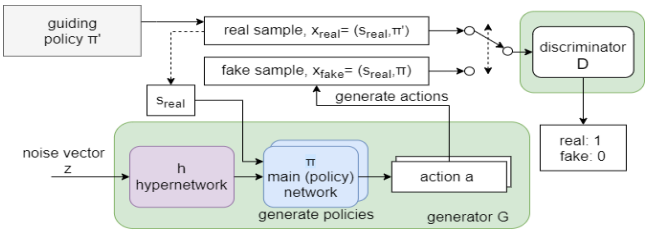

Figure 1: Architecture of learning distribution of policy parameters through hypernetwork

In CEM, policy parameters $\boldsymbol{\phi}$ are sampled from parametric multivariate Gaussian distribution as discussed in Sec 2.1. The parametric distribution approach is known to perform undesirably in high-dimensional problems, due to the severe constraints on the distributions that can be represented. The current limitation lies in the difficulty in learning non-parametric distributions. However, recent advancements in deep learning and generative neural networks [3, 6] have opened up possibility to apply non-parametric distributions in CEM to boost its performance. We propose an adversarially-trained hypernetwork [4] to generate weights for a separate main policy network, representing a more expressive distribution of policy parameters.

The architecture of our proposal is as shown in Fig. 1, analogous to as Generative Adversarial Network (GAN). It consists of a generator G and a discriminator D. A guiding policy is required as an auxiliary learning target. The generator consists of two neural networks, the hypernetwork and the main policy network. The hypernetwork $h(\phi|z;\alpha)$ is conditioned by the Gaussian input noise z along with network parameters α and represents the distribution of policy parameters. $h(\phi|z;\alpha)$ generates the network weights ϕ for the policy network $\pi(a|s;\phi)$, which in turn predicts the action probabilities for a given state s. Experiments show that our proposed approach outperforms its parametric counterpart, with significant difference especially in early training stages.

3.2 Distributions of Returns

Distributional RL [1] is an example of applying distribution concept to the returns component of RL. It is an extension to Q-learning, which the goal is to model the distribution of returns instead of the expectation of returns. The distributional Bellman equation could be expressed analogously to the original Bellman equation:

$$Z^{\pi}(s,a) \overset{dist.}{:=} R(s,a) + \gamma Z^{\pi}\left(S_{t+1}, A_{t+1}\right) \quad (6)$$

Distributional RL has empirically shown strong evidence of good performance compared to its expectation counterpart. It also showed complementary benefit of combining with risk-sensitive applications. One plausible direction is extending its concept to multi-agent scenarios and combine multiple risk-adverse or risk-seeking policies to improve generalization in unseen circumstances.

3.3 Distributions of Actions

SAC is a state-of-the-art method which incorporates an entropy regularization term in the RL objective (compare with Eqn 1):

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left[R\left(s_t, a_t\right) + \alpha \mathcal{H}\left(\pi\left(\cdot \mid s_t\right)\right) \right] \right] \quad (7)$$

The entropy is defined as $\mathcal{H} = -\alpha \log \pi\left(a_{t+1} \mid s_{t+1}\right)$ for a tractable parametric distribution. This has been the limiting factor in SAC, for its policy (or distribution of actions) employing a reparameterized Gaussian distribution. Specifically, a neural network is used to generate the mean and variance of a Gaussian, so that the entropy of the policy has a closed-form solution. To use a non-parametric distribution, an effective method to estimate its entropy without excessive sampling is required. Inspired by the works of Quantile Value Network [2, 8] in Distributional RL, the approach could be explored and extended to learn a quantile function of the action distributions. With a proper entropy estimation, complex multi-modal actor could be used in SAC and its exploration behaviour could be improved in multi-modal problems.

REFERENCES

[1] Marc G Bellemare, Will Dabney, and Rémi Munos. 2017. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*. PMLR, 449–458.

[2] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. 2018. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*. PMLR, 1096–1105.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*. 2672–2680.

[4] David Ha, Andrew M. Dai, and Quoc V. Le. 2017. HyperNetworks. In *International Conference on Learning Representations (ICLR)*.

[5] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *International Conference on Machine Learning (ICML)*. 1861–1870.

[6] Christian Henning, Johannes von Oswald, João Sacramento, Simone Carlo Surace, Jean-Pascal Pfister, and Benjamin F Grewe. 2018. Approximating the predictive distribution via adversarially-trained hypernetworks. In *Bayesian Deep Learning Workshop, Advances in Neural Information Processing Systems (NeurIPS)*.

[7] Shie Mannor, Reuven Y Rubinstein, and Yohai Gat. 2003. The cross entropy method for fast policy search. In *International Conference on Machine Learning (ICML)*. 512–519.

[8] Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. 2019. Fully parameterized quantile function for distributional reinforcement learning. In *Advances in neural information processing systems*. 6193–6202.