Scalable and Parameter-Free Exploration in Deep Reinforcement Learning

Abstract

1 INTRODUCTION

A reinforcement learning environement is modelled as a Markov decision process (MDP) $M = \langle \mathcal{S}, \mathcal{A}, r, P, P_0, \gamma \rangle$, where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the set of available actions. At time $t = 0$ a state s_0 is sampled from the distribution $P_0(\cdot)$. At each timestep an action $a_t \sim \pi(\cdot|s_t)$ is selected and the agent transitions to a new state state $s_{t+1} \sim P(\cdot|s_t, a_t)$. A scalar reward $r_t = r(\cdot|s_t, a_t)$ is observed. As the agent and environement interact in a sequence of time steps, a history of observations $\mathcal{H}_t = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_t, a_t, r_t)$ is collected. The goal is to find a policy $\pi^\star$, such that sampling actions $a \sim \pi^*(\cdot|s)$ maximizes the expected accumulated and discounted future reward, $J^\pi := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$. An efficient agent must be able to learn from the data it collects, but since the data is dependent on the policy, it must also prioritize to explore states and actions that the agent can learn a lot from.

The efficiency of exploration can be measured in regret. The regret of a policy is the difference in the expected reward obtained by following that policy, and the expected reward of following an optimal policy. An learning algorithm's efficiency in exploration can be measured by its cumulative regret over time. One of the simplest reinforcement learning problems is the multi-arm bandit problem. This is an MDP with no state and no transition probabilities. One exploration strategy employed in most Q-learning algorithms to date, ϵ-greedy is provably inefficient, and has a regret bound that grows linearly with time. There are several optimal algorithms for this problem, but perhaps the simplest one is Thompson sampling [Thompson, 1933]. Thompson sampling approximates a posterior distribution of the expected reward for each action. The next action is decided by sampling rewards for each action from the posterior distribution, and selecting the action that gave the highest reward. Bayesian methods been also been shown to behave efficiently with respect to cumulative regret Osband et al. [2016] on general MDPs.

The Q-function, $Q^\pi(s, a)$, is defined as the expected reward of taking action a in state s and then following policy π thereafter: $Q^\pi := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = S, a_0 = A\right]$. The Bellman operator on Q^π is defined as $\mathcal{B}[Q^\pi](s_t, a_t) := \mathbb{E}_{P(s_{t+1}|s_t,a_t)\pi(a_{t+1}|s_t)} \left[r(s_t, a) + \gamma Q^\pi(s_{t+1}, a_{t+1})\right]$. With the Q-function we can define a policy that always picks the action with the highest Q-value in any state. In deep RL, we model Q with a neural network $\hat{Q}_\omega$. One way to facilitate exploration in this policy is to introduce uncertainty in the Q-value function.

Fortunato et al. [2019] introduced NoisyNet for exploration. These are networks with stochastic weights, where each weight $w_{ij}^{(l)}$ has an added pertrubaition sampled from a noise distribution with standard deviation $\sigma_{ij}^{(l)}$. After initializing $\boldsymbol{w}$ and $\boldsymbol{\sigma}$ such that the network has sufficient stochasticity for exploration, both parameters are learned using standard backpropagation. The approach is similar to variational inference schemes such as Bayes by backprop [Blundell et al., 2015] where the weights of a neural network model are assumed to be normally distributed with mean $\mu_{ij}^{(l)}$ and standard deviation $\sigma_{ij}^{(l)}$. The objective of Bayes by backprop, however, is to approximate the posterior distribution $p(\boldsymbol{w}|\mathcal{D})$ in a task with a fixed dataset $\mathcal{D}$. The parameter distribution in NoisyNet does not necessarily converge to an approximate posterior. This means that it does not have the same guarantee on total regret as methods that approximate a posterior over the value functions [Osband et al., 2016]. They do, however, have experimental results which shows that the function does not always converge to a deterministic solution, but it is unclear why the network learns to introduce more noise into the network parameters.

Fortunato et al. [2019] apply NoisyNet to three reinforcement learning algorithms, DQN, Dueling DQN, and A3C, and show improved performance on all of them. Later NoisyNet was used in the Rainbow algorithm [Hessel et al.,

Submitted to the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021). **To be used for reviewing only**.

2017], a combination of six extensions to the DQN algorithms [Fortunato et al., 2019, Bellemare et al., 2017, Wang et al., 2016, van Hasselt et al., 2015, Schaul et al., 2016, Mnih et al., 2016], that shows state-of-the-art performance across 57 Atari games.

In this paper we will introduce an extension to NoisyNet that puts it into a Bayesian context. A limitation of NoisyNet is that the initial uncertainty in the Q-value function is crucial to exploration. If the uncertainty is too high, algorithm will struggle to learn anything, while if the uncertainty is too low, there is nothing incentivizing exploration, and the algorithm will likely be stuck in a poor local minimum.

2 BACKGROUND

2.1 NOISY NETWORKS

In NoisyNet [Fortunato et al., 2019], the Q-function space can be seen as a stochastic process on the probability space $(\Omega, \mathcal{F}, P)$. The stochastic process can simply be written as $\{Q(\boldsymbol{s}, \boldsymbol{a}) : (\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}\}$. For any $\omega \in \Omega$, $Q(\cdot, \cdot, \omega)$ is a sample function mapping $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$. To simplify notation we will denote sample functions as f. NoisyNet, specifically, is a stochastic process parameterized by its weights and their standard deviation. We will define this parameterized function as g_ϕ, and their sampling distribution as $q_\phi \in \mathcal{Q}$. Given a noise vector ξ, we have $f(\boldsymbol{s}, \boldsymbol{a}) = g_\phi(\boldsymbol{a}, \boldsymbol{s}, \xi)$. This means that NoisyNet gives a new policy function for each noise vector ξ. Although this means that NoisyNet can model stochastic policies, Fortunato et al. [2019] point out that for the loss function it is optimising, there always exsit a deterministic optimal policy. They show through emipirical analysis, however, that this does not mean that the policy necessarily disregards the noise and converges to a deterministic policy.

2.2 FUNCTIONAL VARIATIONAL BAYESIAN NEURAL NETWORKS

There are several ways of approximating the posterior distribution of the weights in a neural network with respect to a prior and a dataset. Typically, the dataset $\mathcal{D}$ is static and with datapoints $\boldsymbol{x}$ and labels $\boldsymbol{y}$, and the prior is defined as a distribution over the weights $p(\boldsymbol{w})$ [Rezende et al., 2014, Blundell et al., 2015, Ritter et al., 2018, Maddox et al., 2019]. By defining a prior over the weights, they can use approximate variational inference methods to approximate $p(\boldsymbol{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})}{p(\mathcal{D})}$. The disadvantage of this is that $p(\boldsymbol{w})$ only acts as a regularizer, and is not used to incorporate prior knowledge. Any prior knowledge we might have about the optimizer function would be very difficult to translate into a prior distribution in the weight space.

Functional variational Bayesian neural networks [Sun et al., 2019] is a variational inference method for neural networks that approximates the posterior distribution in function space. This means that our prior will be a distribution over functions, also known as a stochastic process. Sun et al. [2019] show that for two stochastic processes P and Q, the KL-divergence from Q to P is the supremum of the marginal KL-divergences over all finite measurement sets. Let $P_{\mathbf{X}}$ be the marginal distribution of function values at $\mathbf{X} \in \mathcal{X}^n$, then:

$$\mathrm{KL}[Q \| P] = \sup_{n \in \mathbb{N}, \mathbf{X} \in \mathcal{X}^n} \mathrm{KL}\left[Q_{\mathbf{X}} \| P_{\mathbf{X}}\right], \tag{1}$$

Now the stochastic processes are a neural network q_ϕ and a prior p. In a similar manner to NoisyNet, we sample a function f from q by sampling a random noise vector ξ, and then defining $f(\boldsymbol{x}) = g_\phi(\boldsymbol{x}, \xi)$. They further show that the gradient of the KL-divergence for functions evaluated at the measurement set $\mathbf{X}$ is

$$\begin{aligned}\nabla_\phi \mathrm{KL}[q_\phi(\mathbf{f}^{\mathbf{X}}) \| p(\mathbf{f}^{\mathbf{X}})] &= \mathbb{E}_q\left[\nabla_\phi \log q_\phi(\mathbf{f}^{\mathbf{X}})\right] \\ &+ \mathbb{E}_\xi\left[\nabla_\phi \mathbf{f}^{\mathbf{X}}(\nabla_{\mathbf{f}} \log q(\mathbf{f}^{\mathbf{X}}) - \nabla_{\mathbf{f}} \log p(\mathbf{f}^{\mathbf{X}}))\right].\end{aligned} \tag{2}$$

The measurement set $\mathbf{X}$ is created by concatenating the training points $\mathbf{X}^D$ and a set of M points $\mathbf{X}^M$ drawn from some distribution c that has full support in the space of interesting test cases.

The difficult part in (2) is to estimate $\nabla_{\mathbf{f}} \log q(\mathbf{f}^{\mathbf{X}})$ and $\nabla_{\mathbf{f}} \log p(\mathbf{f}^{\mathbf{X}})$. $\nabla_{\mathbf{f}} \log q(\mathbf{f}^{\mathbf{X}})$ is likely intractable, considering q_ϕ is a neural network with stochastic weights. Depending on how we define the prior, however, $\nabla_{\mathbf{f}} \log p(\mathbf{f}^{\mathbf{X}})$ can be easy to compute analytically. To reduce variance in the gradients we define tractable priors in this paper. To estimate the log-density gradient $\nabla_{\mathbf{f}} \log q(\mathbf{f}^{\mathbf{X}})$, they use a spectral Stein gradient estimator [Shi et al., 2018].

3 METHOD

We will present a method based on functional variational Bayesian neural networks [Sun et al., 2019] that allows the following:

- Near-optimal regret
- Incorporate domain knowledge
- Avoid hyperparameter optimization

We would like to approximate the posterior distribution of the Q-value function. This would (1) give near-optimal regret, (2) let us incorporate domain knowledge through an appropriate prior distribution. We can use the functional variational bayesian neural network (FVBNN) [Sun et al., 2019] framework discussed earlier. NoisyNet uses one sample from Q for each optimization step. We will instead use N samples from Q. This lets us add the KL-term from FVBNN to the loss function that encourages stochasticity. The maximization target becomes:

$$\log p(\mathcal{D}_\omega^N|f) - \mathrm{KL}\left[q(\mathbf{f}^{\mathcal{D}_\omega^N}, \mathbf{f}^M) \| p(\mathbf{f}^{\mathcal{D}_\omega^N}, \mathbf{f}^M)\right], \tag{3}$$

where $\mathbf{f}^{\mathcal{D}_\omega^N}$ is f applied to the dataset $\mathcal{D}_\omega^N$, and $\mathbf{f}^M$ is f applied to a set of i.i.d. random points $M = \{m_i \sim c \mid i = 1, \ldots, k\} \subseteq \mathcal{S} \times \mathcal{A}$ with full support. We will let

$$M = \{(s_i, a_j)\ \forall s_i \in \mathcal{D}_\omega^N,\ \forall a_j \in \mathcal{A}\}. \tag{4}$$

Since we are modeling q_ϕ with a Bayesian neural network, and Q_ω with a neural network, we have a bilevel optimization problem. To solve this, we employ a similar strategy to Fellows et al. [2021], where we have a two-timescale gradient update. ϕ and ω are updated using stochastic gradient descent at different timescales to ensure stable convergence.

Algorithm 1 NoisyNet RL

$\mathcal{D} \leftarrow \emptyset$
$s \sim P_0$
Initialize ϕ, ω
while not converged **do**
  $f \sim q_\phi$
  $a \sim \arg\max_a f(s, a)$
  $s' \sim P(\cdot|s, a)$
  $r = r(s', a, s)$
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{s, a, r, s'\}$
  UPDATENOISYNET($\phi, \omega, \mathcal{D}$)
end while

function UPDATENOISYNET($\phi, \omega, \mathcal{D}$)
  $T \sim \mathcal{D}$
  $f \sim q_\phi$
  $b \sim q_\omega$
  $\Delta_\mathcal{L} \leftarrow 0$
  for $\{s, a, r, s'\} \in T$ **do**
    $\boldsymbol{q} \leftarrow \max_a \mathbf{f}(s, a)$
    $\mathcal{F} \leftarrow \mathcal{F} \cup \mathbf{f}(s, a)$
    $G \leftarrow r + \gamma b(s, a)$
    $\Delta_\mathcal{L} \leftarrow \Delta_\mathcal{L} - \frac{1}{|T|} \nabla_\phi \log p(\boldsymbol{q}|\boldsymbol{G})$
  end for
  $\phi \leftarrow \phi + \eta_\phi \Delta_\mathcal{L}$
  $\omega \leftarrow \omega + \eta_\omega(\omega - \phi)$
end function

Algorithm 2 Functional Bayesian RL

$\mathcal{D} \leftarrow \emptyset$
$s \sim P_0$
Initialize ϕ, ω
while not converged **do**
  $f \sim q_\phi$
  $a \sim \arg\max_a f(s, a)$
  $s' \sim P(\cdot|s, a)$
  $r = r(s', a, s)$
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{s, a, r, s'\}$
  UPDATEPOSTERIOR($\phi, \omega, \mathcal{D}$)
end while

function UPDATEPOSTERIOR($\phi, \omega, \mathcal{D}$)
  $T \sim \mathcal{D}$
  $f \leftarrow \{f_i \sim q_\phi \mid i = 1, \ldots, k\}$
  $\mathbf{b} \leftarrow \{b_i \sim q_\omega \mid i = 1, \ldots, k\}$
  $\mathcal{F} \leftarrow \emptyset$
  $\Delta_\mathcal{L} \leftarrow 0$
  for $\{s, a, r, s'\} \in T$ **do**
    $\boldsymbol{q} \leftarrow \max_a \mathbf{f}(s, a)$
    $\mathcal{F} \leftarrow \mathcal{F} \cup \mathbf{f}(s, a)$
    $\boldsymbol{G} \leftarrow \{r + \gamma b_i(s, a) \mid i = 1, \ldots, k\}$
    $\Delta_\mathcal{L} \leftarrow \Delta_\mathcal{L} - \frac{1}{|T|} \nabla_\phi \log p(\boldsymbol{q}|\boldsymbol{G})$
  end for
  $\Delta_{\mathrm{KL}} \leftarrow SSGE(p, \mathcal{F})$
  $\phi \leftarrow \phi + \eta_\phi(\Delta_\mathcal{L} - \lambda \Delta_{\mathrm{KL}})$
  $\omega \leftarrow \omega + \eta_\omega(\omega - \phi)$
end function

4 EXPERIMENTS

Since our method is essentially an extension of NoisyNet Fortunato et al. [2019], all experiments will compare the performance results of these two methods.

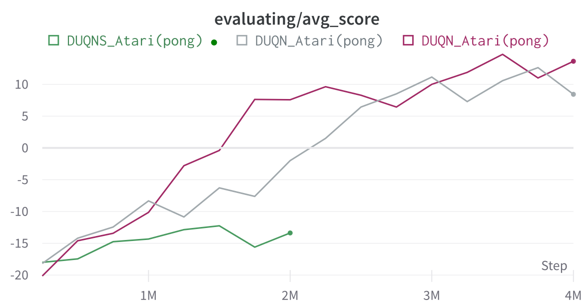

Figure 1: Example figure 1

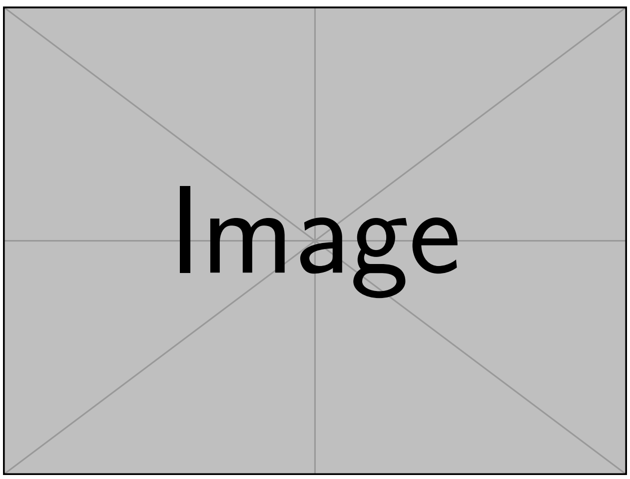

Figure 2: Performance on the CartPole-v1 environment.

4.1 UNINFORMATIVE PRIOR

First, it is interesting to see how our method will compare in environments without the advantage of prior knowledge. We hypothesize that the added KL-term will push the algorithm to keep exploring more than NoisyNet, and

4.2 INFORMATIVE PRIOR

We will also look at how our method can utilize domain knowledge to improve sample efficiency. We will use the following method to calculate the prior distribution on the Q-values:

Discuss how to go from policy prior to Q-value prior.

5 DISCUSSION

References

Marc G. Bellemare, Will Dabney, and Rémi Munos. A Distributional Perspective on Reinforcement Learning. *arXiv:1707.06887 [cs, stat]*, July 2017. URL `http://arxiv.org/abs/1707.06887`. arXiv: 1707.06887.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Network. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, July 2015. PMLR. URL `http://proceedings.mlr.press/v37/blundell15.html`.

Matthew Fellows, Kristian Hartikainen, and Shimon Whiteson. Bayesian Bellman Operators. *arXiv:2106.05012 [cs]*, June 2021. URL `http://arxiv.org/abs/2106.05012`. arXiv: 2106.05012.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Ian Osband, Alex Graves, Vlad Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. Noisy Networks for Exploration. *arXiv:1706.10295 [cs, stat]*, July 2019. URL `http://arxiv.org/abs/1706.10295`. arXiv: 1706.10295.

Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining Improvements in Deep Reinforcement Learning. *arXiv:1710.02298 [cs]*, October 2017. URL `http://arxiv.org/abs/1710.02298`. arXiv: 1710.02298.

Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A Simple Baseline for Bayesian Uncertainty in Deep Learning. February 2019. URL `https://arxiv.org/abs/1902.02476v2`.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous Methods for Deep Reinforcement Learning. *arXiv:1602.01783 [cs]*, June 2016. URL `http://arxiv.org/abs/1602.01783`. arXiv: 1602.01783.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and Exploration via Randomized Value Functions. *arXiv:1402.0635 [cs, stat]*, February 2016. URL `http://arxiv.org/abs/1402.0635`. arXiv: 1402.0635.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *arXiv:1401.4082 [cs, stat]*, May 2014. URL `http://arxiv.org/abs/1401.4082`. arXiv: 1401.4082.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A Scalable Laplace Approximation for Neural Networks. February 2018. URL https://openreview.net/forum?id=Skdvd2xAZ.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized Experience Replay. *arXiv:1511.05952 [cs]*, February 2016. URL http://arxiv.org/abs/1511.05952. arXiv: 1511.05952.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. A Spectral Approach to Gradient Estimation for Implicit Distributions. *arXiv:1806.02925 [cs, stat]*, June 2018. URL http://arxiv.org/abs/1806.02925. arXiv: 1806.02925.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional Variational Bayesian Neural Networks. *arXiv:1903.05779 [cs, stat]*, March 2019. URL http://arxiv.org/abs/1903.05779. arXiv: 1903.05779.

William R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25 (3/4):285–294, 1933. ISSN 0006-3444. doi: 10.2307/2332286. URL https://www.jstor.org/stable/2332286. Publisher: [Oxford University Press, Biometrika Trust].

Hado van Hasselt, Arthur Guez, and David Silver. Deep Reinforcement Learning with Double Q-learning. *arXiv:1509.06461 [cs]*, December 2015. URL http://arxiv.org/abs/1509.06461. arXiv: 1509.06461.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. Dueling Network Architectures for Deep Reinforcement Learning. *arXiv:1511.06581 [cs]*, April 2016. URL http://arxiv.org/abs/1511.06581. arXiv: 1511.06581.