Efficient Exploration with Deep Uncertain Value Networks

Thomas M. Moerland

Department of Computer Science, Delft University of Technology, The Netherlands

INTRODUCTION

The exploration/exploitation trade-off is a core problem of reinforcement learning. Nearly all RL algorithms maintain *point estimates* of the action-value function and/or policy. To ensure exploration, they apply some random perturbation (called 'undirected' exploration) to these point estimates.[1] However, this approach is fundamentally inefficient, as it does not discriminate between an action that has been extensively tried and deemed suboptimal, and an action that has never been tried and requires further exploration (they may both have a low action-value point estimate) [1].

In this work we propose to maintain *distributions* over value functions. Although the action-value is a single number by definition (it is an expectation), it makes sense to treat its estimate as a random variable (from a statistical point of view). The benefit is that the remaining uncertainty about the value of each action provides a natural trade-off between exploration and exploitation, with more exploration as long as we are uncertain, and increasingly more exploitation once we become certain about the environment and task.

TWO TYPES OF VALUE FUNCTION UNCERTAINTY

Uncertainty in the value function may for the RL setting originate from two sources[2][2]

1. **Visitation uncertainty**: when we have infrequently (or even never) visited a particular state-action pair, we should be inherently uncertain about it. This is the traditional statisitical uncertainty that is also studied in the bandit setting.

2. **Bellman uncertainty**: we may visit a certain state-action pair more often, but if we are still uncertain about what to do next, then repeatedly visiting the state-action pair should not make us certain yet about its value. More precisely, for a 1-step bootstrap estimate of the value of state-action we plug in the current value estimate of one-step ahead. However, this value is uncertain itself, because we might not know what is optimal to do from that point. This type of uncertainty makes RL fundamentally different from the bandit setting, as we should effectively propagate uncertainty through the Bellman equation/MDP.

For this work we explore a solution to each type of uncertainty. We focus on function approximation methods based on deep (non-linear) neural networks.

1 For example ϵ-greedy/Gaussian noise for a discrete/continuous action space, respectively.

2 There is a third possible source of uncertainty caused by an uncertain, learned transition model, known as model-based RL. We ignore this problem for this work (we always sample in a ground-truth simulator).

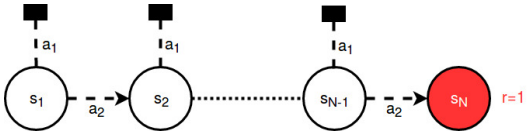

Figure 1: Chain domain, based on Osband et al. [1].

1. **Bayesian drop-out**: Bayesian drop-out [3] in neural networks provides a simple method to approximate/sample from the posterior predictive distribution, which tracks the visitation uncertainty mentioned above.

2. **Gaussian Q-value propagation**: We propagate uncertainty by parametrizing every action-value als a Gaussian, where we learn both the mean μ and standard deviation σ. Then, when we bootstrap the value of an state-action, we sample m values from the next state (s') value distribution, instead of sampling only the mean (once).

As a policy, we then perform Thompson sampling [4] on the uncertain values, which selects each action with probability equal to the probability that it is the optimal one, when averaging out all uncertainty.

EXPERIMENTS

We study these ideas on the Chain domain (Fig. 1). This MDP consists of a chain of states $\{s_1, s_2..s_N\}$ of length N and two actions $\{a_1, a_2\}$. The agent should learn to walk all the way to the end of the chain, by repeatedly taking action a_2, to receive a reward of 1. All actions a_1 terminate the episode without reward. Although the domain has small dimension, it is actually very challenging for 'undirected' exploration methods (as there is no information until we hit the correct chain for the first time, which makes their exploration time scale exponentially).

Fig. 2 shows the results of ϵ-greedy and Thompson sampling on the two types of uncertainty, for different chain lengths N. For the short chain, we see Thompson sampling learning much faster than the ϵ-greedy methods. When we increase the length of the chain, we see that ϵ-greedy methods start to completely fail, while the Thompson sampling methods still solve the problem. Most stable performance is obtained by the Bellman uncertainty propagation, which is further improved when we pre-train the network (after random initialization) on $\mathcal{N}(0, 5)$ noise to remove any initialization bias.

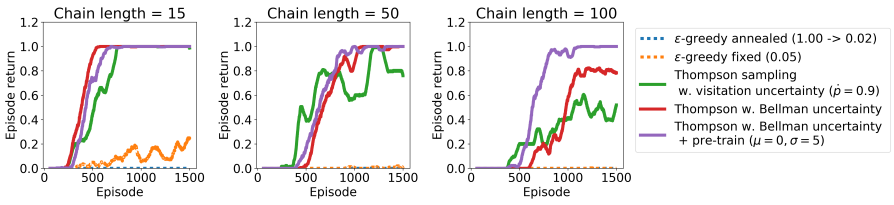

Figure 2: Learning curves for different exploration methods, split-up for different Chain domain lengths. Curves averaged over 5 repetitions.

CONCLUSION

We introduced uncertain value networks, based on Bayesian inference in neural networks, to track distributions over value functions. Our results show a vast increase in exploration time in a domain with a specific exploration challenge. Future work includes investigation in high-dimensional domains as well as joint modelling of both types of uncertainty.

REFERENCES

[1] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*, 2014.

[2] Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian Q-learning. In *AAAI/IAAI*, pages 761–768, 1998.

[3] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[4] 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.