Motivating Exploration in Reinforcement Learning

Bob Wei
Department of Computer Science
University of Waterloo
q25wei@edu.uwaterloo.ca

Akshay Patel
Department of Computer Science
University of Waterloo
akshay.patel@uwaterloo.ca

Samir Alazzam
Department of Computer Science
University of Waterloo
q25wei@edu.uwaterloo.ca

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

Exploration is something that comes naturally to humans, almost as if there exists an innate sense of curiosity craving to experience something new, something yet to be seen. Through observing human behaviour, it is clear that the ability to explore a domain is critical to the process of discovery and learning, regardless of the task being considered.

And yet, even with the massive computational resources available today, the most impressive artificial learning agents struggle tremendously in balancing explicit exploration of the environment and stable convergence towards a useful, learned policy. Without adequate exploration and randomness, the agent can easily become stuck and converge towards a flawed policy early on in learning; the agent's policy reflects high certainty on its understanding of the environment when, in reality, it has yet to explore the majority of the state space. On the other hand, Excessive exploration can lead to unstable rewards and updates throughout learning which are detrimental to learning a useful policy. To make matters worse, there is a huge breadth of algorithms and hand-crafted techniques used in reinforcement learning, many of them requiring method or task specific hyperparameter tuning. This in turn leads to very specialized formulations across the field for encouraging exploration.

In this work, we explore a general formulation of curiosity that aims to motivate exploration in reinforcement learning (RL) agents while remaining invariant to the algorithm and task in question. We explore the effects of including an entropy term (of the policy action probability distribution) and a curiosity module based on that of **cite**, which provides an intrinsic reward signal. Due to the scope of the porject, we focus specifically on the Advantage Actor-Critic (A2C) RL algorithm in the *Pong*, *Seaquest*, and *Breakout* Atari environments. These tasks were chosen due to the differences in mechanics and their respective state spaces, specifically the density and magnitude of the extrinsic reward (i.e. reward received from environment itself).

We compare the performance of the agent in the various environments with and without the mentioned exploration factors. **Add more on the experimental results once that's finalized**

2 Related Work

Someone write this pls

3 Methods

In this section, we present our baseline RL framework using the A2C algorithm which is based on the previous works of **cite**. We then describe the formulations of entropy and curiosity based learning factors to encourage environment exploration.

3.1 Advantage Actor-Critic (A2C)

Our baseline is built around A2C, which is an on-policy learning algorithm. The core decision making of the agent stems from the policy network (π), which is also referred to as the *Actor*. The policy network is learned and its weights (θ_π) are updated via the standard policy gradient equation, wherein the rewards r_t for a trajectory t are weighted by the negative log likelihood of that trajectory. Minimizing this loss $\mathcal{L}_P$ is equivalent to updating π such that the probability of high reward trajectories are maximized.

$$\mathcal{L}_P = \sum_t - \log \pi(s_t; \theta_\pi) r_t \tag{1}$$

A2C also uses the value network (V) or *critic* which predicts the accumulated, discounted rewards R_i over the episode timesteps i. The advantages can be computed as $A_i = R_i - V_i$, where V_i are the predictions from V, and are used in place of the actual rewards in the policy update 1, reducing the variance of $\mathcal{L}_P$, which is a common downfall of on-policy methods. We train the value network V with the update equation 2.

$$\mathcal{L}_v = \frac{1}{n} \sum_i (R_i - V_i)^2 \tag{2}$$

3.2 Motivating Exploration

4 Experiments

4.1 Implementation Details

4.2 Quantitative Results

4.3 Qualitative Results

5 Discussion

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in Section **??** regarding figures, tables, acknowledgments, and references.

6 Figures

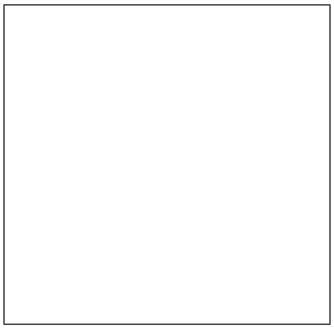

Figure 1: Sample figure caption.

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

7 Tables

Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

References

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.