EXPLORATION POLICIES IN DQN

Sara Silvestrelli
Dipartimento di Economia, Metodi Quantitativi e Strategie di Impresa
Università degli studi di Milano-Bicocca

ABSTRACT

We want the agent to find the best solution as quickly as possible. However, choosing the right action to take too quickly can be counterproductive since usually at the beginning the environment is unknown and must be explored. This leads to the so called exploration-exploitation trade off: the exploration phase looks for more informations of the environment to make better decisions in the future, occasionally taking a random action instead of the supposed optimal one. The exploitation phase maximizes the reward with the known information.

***K*eywords** Artificial Intelligence · DQN · Exploration · Softmax · Boltzmann · Epsilon-Greedy

1 Introduction

Having the agent explore the environment too much can result in missed opportunities because even when an action appears to be the optimal one, wasting both time and compute resources, the agent continues to explore and may miss actions that lead to a better reward. In contrast, not exploring it enough can lead to not identifying new patterns or appropriate strategies. But exploitation is also risky. Exploiting only the information available can lead to the agent being stuck in a local optimum. Not exploiting it enough can have the opposite effect of not acting wisely when necessary.

To avoid missing best rewards with actions that are not observed yet, two possible solutions are the Epsilon-Greedy Policy and the Softmax Policy.

2 Epsilon-Greedy Policy

The $\epsilon-$geedy strategy is included in the original DQN paper [1]. At first the agent does not understand the environment and prefers to explore rather than exploit the information that is nonexistent at the beginning. One solution is to sample at each stage t from a uniform distribution $u_t \sim Unif(0, 1)$. If u_t is below a certain threshold the agent will explore the environment, otherwise it will exploit the already known information.

This means to select the action with the highest estimated future reward and randomly, with small probability ϵ, extract an action to be performed among those possible with equal probability. This action selection rule is called ϵ-greedy method and it improves agent's chances of recognizing the optimal action.

Thus this strategy takes an exploratory action with probability ϵ and a greedy action with probability $1 - \epsilon$.

Article Settings The value of the ϵ was set such that starting from 0.99 it gradually fell deterministically with each episode until it reached a constant minimum value of 0.01 until the end of the training data.

3 Softmax (Boltzmann) Policy

The Softmax method [2] uses a Gibbs, or Boltzmann, distribution. While exploring the agent creates an action distribution that describes how optimal an action is according to the data gathered by the agent.

The original action distribution is defined by the temperature parameter τ which is a positive number that makes the agent choose between picking the action randomly (random policy) and always picking the most optimal action (greedy policy). Softmax utilizes action-selection probabilities that are determined by ranking the value-function estimates using a Boltzmann distribution.

The softmax policy choose action a at the given state s with the probability $\pi(a|s)$ where

$$\pi(a|s) = Pr\{a_t = a|s_t = s\} = \frac{e^{\,Q(s,a)/\tau}}{e^{\,\sum_b Q(s,b)/\tau}}$$

where $T \in (0, \infty)$ is called temperature parameter. High value of the temperatures cause all actions to be nearly equiprobable since there are almost no differences between probabilities and we end up with a random policy. Low value of the temperatures cause greedy action selections since differences between original action probabilities become more substantial and when τ is very close to zero the most probable action is selected all the time.

Article Settings $\tau = 0.5$.

4 Conclusions

Both methods have only one parameter that must be set. The $\epsilon-$greedy method may be useful in continuous state-spaces since no memorization of exploration specific data is required. Although a weakness of this strategy is that it is not clear which value of ϵ leads to better results and this can be a very time-consuming task. This approach is not appropriate in the case certain actions are extremely worse than others. A natural solution is to select random actions with probabilities proportional to their current values defining a state-dependent exploration-probability.

References

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.

[2] Michel Tokic and Günther Palm. Value-difference based exploration: Adaptive control between epsilon-greedy and softmax. In Joscha Bach and Stefan Edelkamp, editors, *KI 2011: Advances in Artificial Intelligence, 34th Annual German Conference on AI, Berlin, Germany, October 4-7,2011. Proceedings*, volume 7006 of *Lecture Notes in Computer Science*, pages 335–346. Springer, 2011.