A Novel Reinforcement Learning Algorithm

Daniel Eftekhari (daniel.eftekhari@mail.utoronto.ca)
Department of Computer Science, Bahen Centre, 40 St. George Street
Toronto, Ontario M5S 2E4 Canada

Abstract

Reinforcement learning has increasingly established itself as a focal point for innovative research in the behavioural sciences, contributing to a better understanding of decision-making in animals, and enabling the development of predictive models of behaviour. However, the mechanism by which animals evaluate the trade-offs in utility between explorative and exploitative policies remains evasive. Despite this, it has become clear that, under varying environmental circumstances, an explorative policy provides the agent with greater utility. In this work, we present a novel learning algorithm with inherently explorative properties. We compare the algorithm's performance to Q-learning, a well-studied algorithm, in a series of experiments with either unchanging or varying environments. Under certain circumstances, our novel learning algorithm is comparable in performance to Q-learning, and often establishes an optimal policy in fewer iterations. However, the novel algorithm's performance under varying environments was somewhat inferior in comparison to Q-learning. Future work will focus on adopting a more biologically plausible discount factor, as well as a modification to the algorithm so that agents can more effectively adapt to changing environments.

Keywords: Reinforcement learning; novel algorithm; exploration vs. exploitation; Q-learning;

Introduction

Reinforcement learning (RL) is an active area of research in the artificial intelligence community. Over the past decade, insights from RL have increasingly been applied to the behavioural sciences. Understanding how animals consider the trade-offs when deciding between an explorative or exploitative decision policy, as well as (at a high level but nevertheless a quantitative one) what dictates animal learning, have been of interest.

As a brief overview, reinforcement learning differs from supervised learning in that although inputs (agent decisions) are attributed to outputs (rewards), there is no one-to-one relationship between action and outcome. Instead, agents must learn, of the sequence of actions taken, which combination led to favourable outcomes and which to unfavourable ones, and subsequently update their decision policy to maximize future utility.

In this work, a novel approach to reinforcement learning with biological plausibility, in which rewards affect policies directly, is investigated, and its performance is contrasted to Q-learning, a well-studied reinforcement learning algorithm. In the next section, the novel learning algorithm and its rationale is presented, followed by a simulation conducted to assess algorithm performance with respect to Q-learning.

Methods and results

Environment

The environment consists of an *l x m* grid, where *l* is the number of rows and *m* the number of columns. Coordinates on the grid are described by (x,y), corresponding to the rows and columns respectively. The grid consists either of free squares, walls or rewards. Agents cannot move onto a square that is occupied by a wall. The environment is *9 x 6* blocks, with rewards between *0* and *1*.

At any time, an agent at position (x,y) can move to positions $(x\text{-}1,y)$, $(x,y\text{-}1)$, $(x+1,y)$ or $(x,y+1)$, assuming these squares are not occupied by a wall. The agent starts each trial at $(\frac{l}{2},0)$. The environment is initialized by making any legal move equally probable at every square in the grid, so that movement starts off as completely random. Once the agent lands on a reward, the iteration is terminated and a new trial begins.

Novel learning rule

Suppose that for any position (x,y), κ is the set of legal actions, and *j* represents the move that was made. The update rule for paths that were not chosen at (x,y) is given by

$$p_i = p_i - \frac{\alpha(1-p_j)R\gamma^s}{(n-1)}$$

where $i \in \kappa$, $i \neq j$ and for the path that was chosen,

$$p_j = p_j + \alpha(1-p_j)R\gamma^s$$

where α is the learning rate, *R* is the reward that was obtained, γ is the discount factor, *s* is the number of moves removed from the reward and *n* is the number of moves that were possible in position (x,y).Note that there is no guarantee that the probabilities will remain positive for $n > 2$. (Perhaps the term probability is thus misleading, unless the next part of the algorithm is considered.) The softmax operation is subsequently applied to the resulting values to ensure all probabilities lie within 0 and 1, i.e.

$$p_u = \frac{\exp(-\tau * p_u)}{\sum_{v=1}^{n} \exp(-\tau * p_v)}$$

where τ is the gain, which governs the exploitation-exploration dichotomy. At the end of each iteration, τ is multiplied by a factor $\zeta > 1$, which causes the algorithm to increasingly shift from exploration to exploitation.

Learning rule rationale

It has been shown that in varying environments (environments where reward locations are changing, or reward values altered), it is beneficial to add noise to the perceived utility of already learned policies, as this allows an agent to once again explore the environment and later exploit the highest-utility option. Conversely, an inherently exploitative learning rule will be unable to adjust to the changing environment quickly, since it will have learned a strong association to the initial rewards. One study describes a model that matches behavioural data from animal studies corresponding to the frontal and locus coeruleus mechanisms. This highlights the notion that in variable environments, a tendency for exploration is important for maximizing utility.

Although the authors in the aforementioned study argue that a learning rule should allow for exploration once an environment is subject to uncertain rewards, we argue that an inherently explorative learning rule, but which allows for exploitation once rewards have been realized, is a more robust approach.

This led us to develop the novel learning algorithm, which has an inherently explorative component, stemming from the term $(1 - p_j)$. Consequently, high probability policies, when chosen, only grow marginally in probability if a reward is obtained, whereas low probability policies, when chosen, grow significantly in probability. This inherently contributes an explorative property - decisions that are unlikely to occur increase significantly in likelihood when a positive reward is associated with them, and decisions highly likely to occur grow only marginally in probability.

Q-Learning

Q-learning is a well-studied reinforcement learning algorithm which, despite its simplicity, is effective in identifying optimal policy rules. Much like our novel algorithm, Q-learning does not require *a priori* knowledge of its environment to work. Additionally, Q-learning is guaranteed to find the optimal policy for any Markov decision process (MDP). The reader is encouraged to visit any textbook on RL for a more comprehensive review of Q-learning.

Simulation: Overcoming local optima

Throughout our experiments, we observed the following trends. Small changes to τ and ζ have the most drastic effects on learning - high values cause strictly exploitative behaviour, while low values cause strictly explorative behaviour. The task is to strike a balance between the two extremes. A high learning rate (high α value) attributes high importance to almost any outcome, causing probabilities to change drastically from iteration to iteration. On the other hand, a low α does not attribute a large importance to any particular iteration, and instead the probability converges slowly over time to the optimal policy. This explains why a small α value lead to better optimal policies in our experiments, although a larger number of iterations is required for convergence.

Figure 1 shows an environment with two rewards. The distance to the orange block is 2 moves, while the distance to the red block is 6 moves. The question is, will the agent overcome the nearby local optima and learn a policy which corresponds to the global optimum? Using default hyperparameters, Figure 2 demonstrates what percentage of the time the algorithms optimal policy leads to the global optimum (for the novel learning algorithm and for Q-learning) after a given number of iterations.

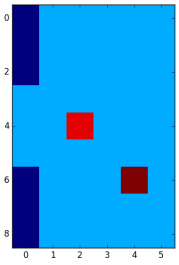

Figure 1: A local optima (orange: value 0.1, distance 2 away) and a global optimum (red: value 0.4, distance 6 away).

The novel learning algorithm is able to reach a high level of performance with fewer iterations than Q-learning. This is because of its inherently explorative behaviour, allowing it to learn its environment very well, and only exploit the resource that is likely to be the optimal one. This is further supported by the sudden transition between relatively poor and excellent performance in the novel learning algorithm at around 8500 iterations. The algorithm takes the first 8500 iteration to explore the environment, and then (due to exponential growth in τ) makes a firm decision on which reward to pursue.

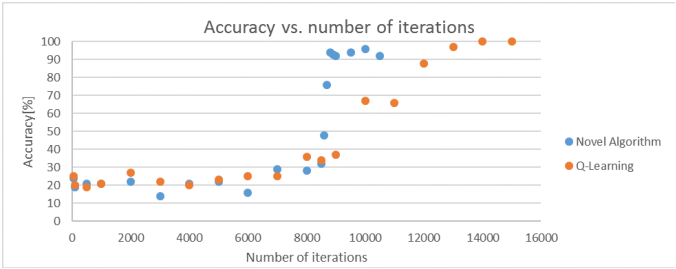

Figure 2: Accuracy vs. number of iterations for novel learning algorithm and Q-learning.

Future steps

In contrast to this work, empirical research suggests most animals use hyperbolic discounting for future rewards. However, there is a dynamic inconsistency inherent in hyperbolic discounting, which leads to choices which are inconsistent over time, and thus lead to erroneous decision-making when predicting the value of discounted rewards. We therefore aim to assess whether, in our experiments, exponential discounting would indeed perform better than hyperbolic discounting, as the literature suggests it should.