Title

Pedro Almeida[1], Siddhartha Verma [1,2]†

[1]Department of Ocean and Mechanical Engineering, Florida Atlantic University, Boca Raton, FL 33431, USA

[2]Harbor Branch Oceanographic Institute, Florida Atlantic University, Fort Pierce, FL 34946, USA

(Received xx; revised xx; accepted xx)

Test

Key words: Authors should not enter keywords on the manuscript, as these must be chosen by the author during the online submission process and will then be added during the typesetting process (see http://journals.cambridge.org/data/relatedlink/jfm-keywords.pdf for the full list)

1. Introduction

Reinforcement Learning (RL) is an area of machine learning where an agent learns optimal behavior through repeated interactions with an environment that maximize some notion of a cumulative reward.

Why it differs from other methods of machine learning

Some common applications (famous examples).

Challenges in designing a system

Introduction of basic concepts.

Markov decision process (MDP):

2. Methods

2.1. *Q-Learning*

2.1.1. *Theory of Q-Learning*

Q-Learning is a model-free RL algorithm whereby an agent learns the value of an action for a given state by calculating the expected reward for an action taken in a given state. Originally proposed by Watkins in 1989 Watkins (1989),

- Works well for discretized environements (like grids)
- Can work with continuous with binning
- Becomes more difficult when size of problem increases (space and time complexity explode)
- Exploring all states in q-learning often takes too long

2.1.2. *Grid World*

- Grid like 8x8
- One thief (seeking)

† Email address for correspondence: vermas@fau.edu

- Police Officer (Obstacle)
- Gold (Goal)

$$Q^{new}(s_t, a_t) = Q(s_t, a_t) + \alpha \left(r_t + \gamma Q(s_{t+1}, a) - Q(s_t, a_t) \right) \tag{2.1}$$

2.2. *Deep Q-Learning*

2.2.1. *Deep Q-Learning Theory*

A bit better for continuous state domains

2.2.2. *Environment*

Studies (2007)
Brockman *et al.* (2016)

$$\ddot{\theta} = \frac{gsin(\theta) - cos(\theta)\left(\frac{-F - m_p L \dot{\theta}^2 sin(\theta)}{m_t}\right)}{L * \left[\frac{4}{3} - \frac{m_p cos^2(\theta)}{m_t}\right]} \tag{2.2}$$

$$\ddot{x} = \frac{F + m_p L\left(\dot{\theta}^2 sin(\theta) - \ddot{\theta} cos(\theta)\right)}{m_t} \tag{2.3}$$

3. Conclusions

REFERENCES

Brockman, Greg, Cheung, Vicki, Pettersson, Ludwig, Schneider, Jonas, Schulman, John, Tang, Jie & Zaremba, Wojciech 2016 Openai gym, arXiv: arXiv:1606.01540.

Studies, Neural 2007 Correct equations for the dynamics of the cart-pole system. *Romania* pp. 1–6.

Watkins, Christopher J C H 1989 Learning from delayed rewards. PhD thesis.