Master thesis on Intelligent Interactive Systems

Universitat Pompeu Fabra

Reinforcement Learning with Options in semi Markov Decision Processes

Sayan Goswami

Supervisor: Anders Jonsson

Department of Information and Communication Technologies, Universitat Pompeu Fabra

Co-Supervisor: M. Sadegh Talebi

Department of Computer Science, University of Copenhagen

September 2021

Master thesis on Intelligent Interactive Systems

Universitat Pompeu Fabra

Reinforcement Learning with Options in semi Markov Decision Processes

Sayan Goswami

Supervisor: Anders Jonsson

Department of Information and Communication Technologies, Universitat Pompeu Fabra

Co-Supervisor: M. Sadegh Talebi

Department of Computer Science, University of Copenhagen

September 2021

Universitat
Pompeu Fabra
Barcelona

Contents

Dedication

(Optional, if used placed on a right page next to an empty left page)

I would like to dedicate this work to...

Acknowledgement

(Optional, if used placed on a right page next to an empty left page)

I would like to express my sincere gratitude to:

- My supervisor
- My co-supervisor
- My family

Abstract

The abstract should have at least 200 but not more than 600 words. Placed on a right page next to a blank left page. A list of keywords (approximately 3 to 5) should be just below the abstract, preceded by the word "Keywords". Keywords should be separated by ";".

Keywords: Imaging techniques; Cloud computing; Alzheimer

Chapter 1

Introduction

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long

1.1 Motivation

1.2 Objectives

1.3 Structure of the Report

Chapter 2

The Reinforcement Learning Paradigm

Reinforcement Learning (RL) is a sub-field of machine learning wherein an agent learns to take optimal decision in an unknown environment by trial and error. RL differs from the other branches of machine learning – namely supervised learning and unsupervised learning. Supervised learning, involves learning from labelled data whereas unsupervised learning, involves finding the underlying structure of unlabelled data. RL involves an agent actively exploring an interactive environment sequentially and performing actions to maximize a "reward" metric over a number of steps.

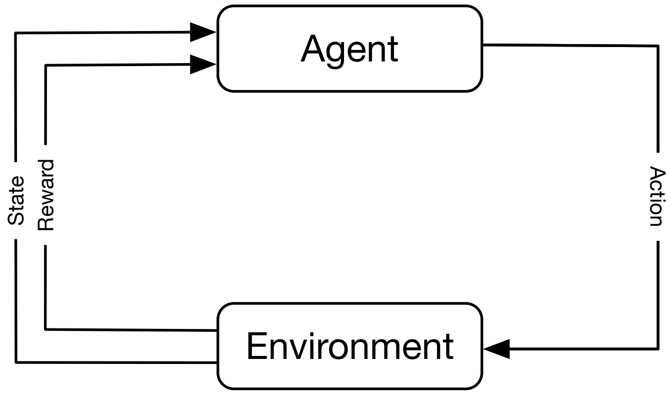

Figure 1: The basic reinforcement learning scheme.

In the reinforcement learning paradigm, at each time step, an agent interacts with the environment by choosing an action based on its present state. Then the agent proceeds onto a new state as dictated by the dynamics of the environment. It receives a scalar feedback, a reward, pertaining to its previous action in the process.

2.1 Basic Terminologies in RL

In this section, we introduce the basic terminologies that will bes used through the course of this manuscript to convey ideas related to reinforcement learning.

2.1.1 Agent

Reinforcement learning is a sequential decision making problem wherein at each time step a entity takes decision based on it's current and past observations. This entity is the agent.

2.1.2 Environment

The agent takes decisions at each time step to influence it's surroundings, the environment. In turn, the agent receives a scalar feedback signal from the environment.

2.1.3 State

A state is a representation of the agent's environment. State can be segregated into two components – the environment state (S_t^e) and the agent state (S_t^a). S_t^e is the environment's private representation, it is inaccessible by the agent. S_t^a is the agent's internal representation. It is used by the learning algorithms. Hereon, we shall use S_t to denote to the agent's internal state representation.

2.1.4 Action

It is a set of possible moves available to the agent at any state. Taking an action may cause the agent to transition from one state to another.

2.1.5 Policy

A policy ($\pi : \mathcal{S} \mapsto \mathcal{A}$) is a mapping from the agent's state representation to the set of actions. A policy may be deterministic, meaning each state maps to a single action or stochastic, meaning each action induces a probability distribution over a set of possible actions, in nature.

2.1.6 Reward

It is a scalar feedback received from the environment as a consequence of the agent's action. A characteristic of reinforcement learning is the delay of rewards. Consequence of an action is not immediate, it's effect is observed in the future.

Discounted Reward Setting

In the discount reward setting at timestep t, the return G_t is defined as the discounted sum of the rewards obtained starting from timestep $t+1$. In other words,

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

where, $\gamma \in [0, 1]$ is the discount rate and T is the time horizon.

Starting from state s at time t and then following policy π, the value of the state s is given by

$$V^{\pi}(s) = E_{\pi}\left[G_t \mid S_t = s\right]$$

The discounted reward setting is suitable for finite as well as infinite time horizon tasks. However, discounted rewards are necessary only in cases where the cumulative reward is unbounded, that is, in the case of infinite time horizon tasks. In finite horizon tasks, simply the cumulative sum of rewards can be used.

Average Reward Setting

The average reward setting is more suitable for cyclical tasks. The first average reward RL, R-Learning method was introduced by Schwartz [1]. Average reward is defined as a time average of the cumulative reward. In other words, the gain (average reward) $\rho^\pi(s)$ for state s, under policy π is given as

$$\rho^\pi(s) = \lim_{T\to\infty} \frac{\sum_{t=0}^{T-1} R_t(s)}{T}$$

where, $R_t(s)$ is the reward received at time t starting from state s and thereafter following policy π.

2.2 Markov Processes

A Markov Process is a random process defined by the 2-tuple $\langle \mathcal{S}, \mathcal{P} \rangle$. Set $\mathcal{S}$ is the set of states (maybe infinite, considered finite in our work). $\mathcal{P}$ is the transition probability kernel such that

$$\mathcal{P}_{ss'} = P\left[S_{t+1} = s' \mid S_t = s\right]$$

where $s, s' \in \mathcal{S}$.

2.3 Markov Reward Processes

A Markov Reward Process builds up on top of a Markov Process. It is represented by a 3-tuple $\langle \mathcal{S}, \mathcal{P}, \mathcal{R} \rangle$. $\mathcal{R}$ is the reward kernel such that,

$$\mathcal{R}_t(s) = E[R_{t+1} \mid S_t = s]$$

where $s \in \mathcal{S}$.

2.4 Markov Decision Processes

Markov Decision Processes (MDPs) builds up on top of a Markov Reward Process. It is represented by a 4-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$. Set $\mathcal{A}$ is the set of actions (maybe infinite, considered finite in our work).

2.4.1 Communicating States

Two states s_1 and s_2 in an MDP are said to be communicating under a policy π if there is a non-zero probability of reaching each state from the other in zero or more state transitions.

2.4.2 Recurrent States

A state s is recurrent if there is a non-zero probability of returning to the same state under a policy π in zero or more state transitions. In other words, the probability of eventually returning to state s is unity.

2.4.3 Ergodic Class of States

A subset of states are said to belong to an ergodic class if they are recurrent, communicate with each other and do not communicate with any states outside the class. A Markov Chain is irreducible if $\mathcal{S}$ forms an ergodic class.

2.4.4 Ergodic MDP

An MDP is ergodic if the transition matrix corresponding to every policy has a single recurrent class of states.

2.4.5 Communicating MDP

An MDP is communicating if each pair of states communicate with each other under some stationary policy π.

2.5 Options Framework

First introduced by Sutton et al. [2], options are a means to represent temporally extended actions in an MDp. An option consists of three components: (a) an option policy $\pi_o : \mathcal{S} \mapsto \mathcal{A}$ (b) an initiation set $\mathcal{I} \subset \mathcal{S}$ and, (c) a termination condition $\beta : \mathcal{S} \mapsto [0, 1]$.

An option $o = \langle \mathcal{I}, \pi_o, \beta \rangle$ is available in a state s iff $s \in \mathcal{I}$. If an option is executed, actions are chosen according to the option's policy π_o till the option terminates. An option terminates stochastically based on β, its termination condition.

TODO: Add why options are required. Describe Markov, semi-Markov options.

2.6 Semi Markov Decision Processes

Semi Markov Decision Processes (SMDPs) build up on top of MDPs with Options. SMDPs are special kinds of MDPs that are suitable for modelling continuous time discrete event system. Actions in SMDPs take varying amounts of time to terminate. The SMDP framework is intended to model temporally extended courses of actions.

The key constituents of a SMDP are (a) a set of states (b) a set of options (c) a transition probability kernel condition on the options (d) a well defined set of rewards for ever state option pair .

Chapter 3

Learning Algorithms

3.1 UCB

3.2 UCRL

3.3 EVI

3.4 Our Contribution

Chapter 4

Experimentation

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

4.1 Environment

4.1.1 River Swim Domain

4.1.2 Grid World Domain

4.2 Experimentation

Chapter 5

Conclusion

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

List of Figures

List of Tables

Bibliography

[1] Schwartz, A. A Reinforcement Learning Method for Maximizing Undiscounted Rewards. *Machine Learning* **Proceedings of the Tenth International Conference** (1993). URL `https://antonjazz.com/x/grab/AntonSchwartzReinforcementLearningML93.pdf`.

[2] Sutton, R. S., Precup, D. & Singh, S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* **112**, 181–211 (1999). URL `https://linkinghub.elsevier.com/retrieve/pii/S0004370299000521`.

Appendix A

First Appendix

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

Appendix B

Second Appendix