

BSc Hons: Computer Science

Literature Review

Reinforcement Learning

Goal Selection Strategies for Learning Goal-Oriented Value
Functions



Mamello Seboholi
1851317

Supervised by Steven James

Chapter 1

Introduction

The aim of this project is to extend the Goal-oriented Q-learning algorithm used in logical composition reinforcement learning (RL) by considering different use cases.

Chapter 2

Background and Related Work

2.1 Introduction

This chapter serves to present our findings in related work to this paper. The paper has the following layout: section 2.2 explores knowledge transfer through composition with the focus on concurrent composition; and section 2.4 will serve as a conclusion for this chapter.

2.2 Composition

It is essential that we discuss the idea of *composition* [Todorov 2009] within the context of this project as we aim to improve on the literature focused on this topic.

Nangue Tasse *et al.* [2020] discusses the need for reinforcement learning (RL) agents to have an ability of knowledge transfer as reinforcement learning (RL) problems become expensive. The paper models tasks using Markov Decision Processes (MDPs) where an MDP is defined as a quadruple (4-tuple) composed of the (i) state space S , (ii) action space A , (iii) Markov kernel defined by ρ that takes $S \times A$ to S , and (iv) reward function r that has real values and bounded by a minimum and maximum r values.

RL agent's goal is to work out a policy π that maps S to A , which solves a given task optimally. Extended reward function and extended Q-value functions are defined. Tasks, as well as the extended Q-value functions are defined as Boolean algebra.

This paper, along with Todorov [2009], Saxe *et al.* [2017], Haarnoja *et al.* [2018], Van Niekerk *et al.* [2019], Hunt *et al.* [2019], and Peng *et al.* [2019] focus on concurrent composition, where novel tasks are formed by combining previously learned tasks. Another group of literature focus on sequentially chaining learned policies to solve complex tasks, examples of this is options [Sutton *et al.* 1999] and hierarchical reinforcement learning [Barto and Mahadevan 2003].

This paper is the basis for this project with the focus being on how agents choose between maximizing goals and discovering new goals in the environment, which is currently done in a greedy fashion.

2.3 Multi-Armed Bandits

The balance between exploration and exploitation is a well known problem in RL. and much of this project is focused on idea. It has been researched rigorously with respect to finding the optimal policy for actions, however, there is no significant work relating to the balance between exploration and exploitation for goal-oriented RL.

The literature is usually classified into two types of methods. One type is *undirected* methods, where agents resolve the exploration-exploitation dilemma by using the Q-values, these types of methods seem to perform very well with small to medium problem sizes but do not seem to find optimal policies when the problem scales significantly. Another type is *directed* methods, where agents resolve the exploration-exploitation dilemma by using knowledge about exploration, these methods deal well with increased scale of problems, however, this comes with considerable computation requirements.

Epsilon greedy (ϵ -greedy) is a very simple and popular method used to balance between exploration and exploitation. It is an example of an *undirected* method. It explores the environment with a probability of ϵ , and chooses an action giving the highest reward with a probability of $1 - \epsilon$, where ϵ is chosen to be a very small number within the open interval $(0, 1)$ (Note: An ϵ of 0 means exploitation only and an ϵ of 1 means exploration only). It's origins are not clear, however, it has been used as far back as [Watkins \[1989\]](#) and [Sutton et al. \[1998\]](#). Papers like [Tokic and Palm \[2011\]](#) and [dos Santos Mignon and da Rocha \[2017\]](#) have extended the ϵ -greedy method with an attempt to control the exploration rate (ϵ). An issue with ϵ -greedy is that, during exploration, it chooses equally among the other actions, this is not good when there exists actions with negative rewards.

Another example of an *undirected* method that is also known as a *probability matching* method is *softmax*. It utilizes the Boltzmann distribution (also known as the Gibbs distribution) to encourage random action selection. As with the case of ϵ -greedy method, a user also controls a parameter in *softmax*. The parameter τ known as the *temperature* is a value in the half-open interval $(0, \infty]$ where high values of τ encourage exploration only and values close to 0 encourage exploitation only. The origin of *softmax* method is also not clear, however, it has been used as far back as [Luce \[1959\]](#). [Cesa-Bianchi and Fischer \[1998\]](#) introduces *softmix* which extends *softmax* by changing the way in which the temperature decreases, it uses $\log b/t$ as oppose to $1/t$. [Auer et al. \[1995\]](#) introduced a complicated variant of *softmax* called *Exp3* ("exponential-weight algorithm for exploration and exploitation") and slight variations thereof are introduced in [Auer et al. \[2002b\]](#).

[Vermorel and Mohri \[2005\]](#) does an empirical evaluation of the *undirected* methods and conclude that simple methods like ϵ -greedy tend to perform well than the complicated counterparts in simple enough environments.

[Wyatt \[1998\]](#) introduces Q-value sampling which is a *directed* method also known as a *stochastic* method where the rewards are represented by a probability distribution. The probability distribution takes into account both exploitation (expected reward) and exploration (how uncertain it is for actual reward). An agent then takes an action based on this probability distribution. An issue with this method is that it requires vast amount of data to build the probability distribution. [Dearden et al. \[1998\]](#) uses Q-value sampling along with Myopic value of imperfection [Howard \[1966\]](#) to present a Bayesian based Q-learning method where actions also depend on probability distribution. Thompson Sampling [[Thompson 1933](#)] is another sampling based algorithm.

Upper Confidence Interval (UCB) and specifically UCB1 [Auer *et al.* 2002a] is a stochastic method that is based on optimism. In this method, data is gathered and then used to assign a weight to each arm in the multi-armed bandit problem, this weight is known as the upper Confidence bound. UCB methods move focus from exploration to exploitation, $\log t / N_t(a)$ is used encourage exploration of the environment as $N_t(a)$ remains small for actions that have not been explored for a long time, a parameter c is also used to control level of exploration.

Abed-alguni [2018] uses an optimization algorithm known as Cuckoo Search [Yang and Deb 2009] to balance between exploration and exploitation and calls this Cuckoo Action Selection (CAS). This paper has a different take in that it uses nature, the inspiration is from breeding behavior of some parasitic cuckoo species and how they lay their eggs.

2.4 Conclusion

This chapter serves to showcase literature in Reinforcement Learning and how they have an effect on our research project. These papers will help in determining an appropriate method that will balance between exploration and exploitation in the Goal-oriented Q-learning algorithm for the boolean algebra tasks.

References

- [Abed-alguni 2018] Bilal H Abed-alguni. Action-selection method for reinforcement learning based on cuckoo search algorithm. *Arabian Journal for Science and Engineering*, 43(12):6771–6785, 2018.
- [Auer *et al.* 1995] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pages 322–331. IEEE, 1995.
- [Auer *et al.* 2002a] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.
- [Auer *et al.* 2002b] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.
- [Barto and Mahadevan 2003] Andrew G Barto and Sridhar Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, 13(1):41–77, 2003.
- [Cesa-Bianchi and Fischer 1998] Nicolo Cesa-Bianchi and Paul Fischer. Finite-time regret bounds for the multiarmed bandit problem. In *ICML*, volume 98, pages 100–108. Citeseer, 1998.
- [Dearden *et al.* 1998] Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian q-learning. In *Aaai/iaai*, pages 761–768, 1998.
- [dos Santos Mignon and da Rocha 2017] Alexandre dos Santos Mignon and Ricardo Luis de Azevedo da Rocha. An adaptive implementation of ϵ -greedy in reinforcement learning. *Procedia Computer Science*, 109:1146–1151, 2017.
- [Haarnoja *et al.* 2018] Tuomas Haarnoja, Vitchyr Pong, Aurick Zhou, Murtaza Dalal, Pieter Abbeel, and Sergey Levine. Composable deep reinforcement learning for robotic manipulation. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 6244–6251. IEEE, 2018.
- [Howard 1966] Ronald A Howard. Information value theory. *IEEE Transactions on systems science and cybernetics*, 2(1):22–26, 1966.
- [Hunt *et al.* 2019] Jonathan Hunt, Andre Barreto, Timothy Lillicrap, and Nicolas Heess. Composing entropic policies using divergence correction. In *International Conference on Machine Learning*, pages 2911–2920. PMLR, 2019.
- [Luce 1959] R Luce. *D., Individual Choice Behavior*, 1959.
- [Nangue Tasse *et al.* 2020] Geraud Nangue Tasse, Steven James, and Benjamin Rosman. A

- boolean task algebra for reinforcement learning. *Advances in Neural Information Processing Systems*, 33:9497–9507, 2020.
- [Peng *et al.* 2019] Xue Bin Peng, Michael Chang, Grace Zhang, Pieter Abbeel, and Sergey Levine. Mpc: Learning composable hierarchical control with multiplicative compositional policies. *Advances in Neural Information Processing Systems*, 32, 2019.
- [Saxe *et al.* 2017] Andrew M Saxe, Adam C Earle, and Benjamin Rosman. Hierarchy through composition with multitask lmdps. In *International Conference on Machine Learning*, pages 3017–3026. PMLR, 2017.
- [Sutton *et al.* 1998] Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*. Vol. 135, 1998.
- [Sutton *et al.* 1999] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.
- [Thompson 1933] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.
- [Todorov 2009] Emanuel Todorov. Compositionality of optimal control laws. *Advances in neural information processing systems*, 22, 2009.
- [Tokic and Palm 2011] Michel Tokic and Günther Palm. Value-difference based exploration: adaptive control between epsilon-greedy and softmax. In *Annual conference on artificial intelligence*, pages 335–346. Springer, 2011.
- [Van Niekerk *et al.* 2019] Benjamin Van Niekerk, Steven James, Adam Earle, and Benjamin Rosman. Composing value functions in reinforcement learning. In *International Conference on Machine Learning*, pages 6401–6409. PMLR, 2019.
- [Vermorel and Mohri 2005] Joannes Vermorel and Mehryar Mohri. Multi-armed bandit algorithms and empirical evaluation. In *European conference on machine learning*, pages 437–448. Springer, 2005.
- [Watkins 1989] Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.
- [Wyatt 1998] Jeremy Wyatt. Exploration and inference in learning from reinforcement. 1998.
- [Yang and Deb 2009] Xin-She Yang and Suash Deb. Cuckoo search via lévy flights. In *2009 World congress on nature & biologically inspired computing (NaBIC)*, pages 210–214. Ieee, 2009.



Declaration 2022

I, Seboholi Mamello Justice, (Student number: 1851317)

am a student registered for Introduction to Research Methods in 2022.

This declaration applies to the Literature Review document of
Introduction to Research Methods.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand, Johannesburg may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature: 

Date: 01/05/2022