School of Computing
FACULTY OF ENGINEERING AND
PHYSICAL SCIENCES

Final Report

Planning to Explore in Reinforcement Learning

Jared William Swift

Submitted in accordance with the requirements for the degree of
BSc Computer Science with Artificial Intelligence (Ind)

2022/23

COMP3931 Individual Project

The candidate confirms that the following have been submitted.

Items	Format	Recipient(s) and Date
Final Report	PDF file	Uploaded to Minerva (DD/MM/YY)
<Example> Scanned participant consent forms	PDF file / file archive	Uploaded to Minerva (DD/MM/YY)
<Example> Link to online code repository	URL	Sent to supervisor and assessor (DD/MM/YY)
<Example> User manuals	PDF file	Sent to client and supervisor (DD/MM/YY)

The candidate confirms that the work submitted is their own and the appropriate credit has been given where reference has been made to the work of others.

I understand that failure to attribute material which is obtained from another source may be considered as plagiarism.

(Signature of Student) _____

Summary

<Concise statement of the problem you intended to solve and main achievements (no more than one A4 page)>

Acknowledgements

<The page should contain any acknowledgements to those who have assisted with your work. Where you have worked as part of a team, you should, where appropriate, reference to any contribution made by other to the project.>

Note that it is not acceptable to solicit assistance on 'proof reading' which is defined as the "the systematic checking and identification of errors in spelling, punctuation, grammar and sentence construction, formatting and layout in the test"; see
`https://www.leeds.ac.uk/secretariat/documents/proof_reading_policy.pdf`

Contents

Chapter 1

Introduction and Background Research

1.1 Introduction

- Reinforcement Learning (RL) is based on the concept of learning through experience - through trial-and-error.
- This is akin to the way in which humans and animals learn new skills (link to psychology).
- Motivate need for exploration
- Exploration is a widely studied topic in RL, as it is necessary for learning. Although exploration methods exist based on optimism, cite here, intrinsic motivation, cite here, human interaction, cite here and information theory, cite here, in practice, random exploration is ubiquitous. Randomness precludes efficiency and prohibits explainability.
- In contrast to RL, Automated Planning uses embedded knowledge of the environment, in the form of a model, to determine the optimal sequence of successive actions required to reach a goal. However, Planning relies the model to accurately represent the environment - this cannot be guaranteed and therefore Planning can be quite fragile.
- Combinations of Planning and RL have been developed in order to overcome the fragility of Planning and reduce exploration required for learning, such as DARLING [1], where exploration is constrained to seemingly optimal states by reasoning on the model.
- Discuss issues with current state of the art.
- Within this work we explore the development of a framework that synergises planning and learning in order to drive and constrain exploration by making intelligent hypotheses about the environment, informed by the inherently inaccurate, but still useful, model, previous experience and environmental observations, rather than through randomness. The ultimate goal of this work is to mitigate the effect of the inherent inaccuracies in the model on the quality of learned behaviour; resulting in agents that can learn beyond the inaccuracies of the model, through intelligent exploration.

Reinforcement Learning (RL) is based on the concept of learning through experience - it is a form of trial-and-error learning. An agent learns how to behave in an environment by interacting with it and receiving reinforcement (both positive and negative) through a numerical signal called the reward. RL is grounded in psychology

This is akin to the way in which humans and animals learn new skills

1.2 Reinforcement Learning

Reinforcement Learning (RL) does not fall into either of the traditional machine learning paradigms (supervised and unsupervised learning) - it is a machine learning paradigm of its

own. Within an RL problem a goal-directed decision-making agent learns how to behave in an environment (which may be stochastic). The agent learns by interacting with the environment through actions and observing the affects through its new state and a numerical reward signal. The goal of the agent is to learn how to map states to actions in order to maximise the cumulative long-term reward signal [2]. A Model-Free RL agent has 3 elements:

- **Policy**
- **Reward**
- **Value**

A Model-Based RL agent has the same elements as that in the Model-Free setting, but additionally it has a Model. The model may be learnt by the agent in order to plan on, or it may be given to the agent to inform exploration. Within the Introduction we motivated the need for exploration, however we did not mention a key problem in RL; the *exploration-exploitation trade-off*. The agent needs to explore in order to gain experience and learn, but it must also exploit its learned knowledge - the problem is knowing when to explore and when to exploit.

1.2.1 Markov Decision Processes

The Markov Property states that the future is conditionally independent of the past given the present. An RL problem that satisfies the Markov Property is known as a Markov Decision Problem, and can be modelled by a Markov Decision Process (MDP). MDPs can be fully (MDP) or partially observable (POMDP). Consider an agent interacting with a stochastic gridworld, the agent can easily observe its state, whereas a robot navigating a maze may not be able to observe its exact state, due to uncertainty in sensors, joint readings, etc. In fact, the real-world is a POMDP. Within this work, as a simplification, we assume that the agent is fully able to observe its state. Furthermore that the environment can be discretised (rather than being modelled in a continuous nature); hence we consider **finite** MDPs. A finite MDP is a 4-tuple: MDP $= (S, A, T, R)$ where:

- S is a finite set of states.
- A is a finite set of actions.
- $T : S \times A \times S \to [0, 1]$ is the transition function, which determines the probability of transitioning from a state $s \in S$ to $s' \in S$ with an action $a \in A$.
- $R : S \times A \times S \to \mathbb{R}$ is the reward function, which determines the reward signal, $r \in \mathbb{R}$ received by the agent from transitioning from a state $s \in S$ $s' \in S$ with the action $a \in A$. This reward is extrinsic to the agent; it comes from the environment.

The transition function, T is a key indicator about the nature of the environment. If $\forall s, s' \in S, \forall a \in A$, $T(s, a,' s) \in \{0, 1\}$, then the environment is deterministic, otherwise it is probabilistic. A deterministic environment is one which there is no variance in the outcomes of action in a given state; taking the same action in the same state always produces the same

outcome. Whereas a probabilistic (or non-deterministic) environment has uncertainty associated with transitions.

The Bellman Equation determines the expected reward for being in a state $s \in S$ and following a fixed policy π:

$$V^{\pi}(s) = R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V^{\pi}(s')$$

Where V^{π} is the value function of the policy, π, $0 \leq \gamma \leq 1$ is the discount factor. The Bellman Optimality equation determines the reward for taking the action giving the highest expected return.

$$V^{\pi*}(s) = \text{argmax}_a R(s, a) + \gamma \sum_{s'} P(s'|s, a) V^{\pi*}(s')$$

Where $V^{\pi*}$ is the value function of the optimal policy, $\pi*$, $0 \leq \gamma \leq 1$ is the discount factor.

1.2.2 Model-Free RL

1.2.3 Model-Based RL

1.2.4 Dynamic Programming

- Given a perfect model of the environment, embedded in a MDP, an optimal policy can be computed using Dynamic Programming. However, this assumption of a perfect model is flawed.

1.2.4.1 Policy Iteration

1.2.4.2 Value Iteration

1.2.5 Temporal Difference Learning

- The (temporal) credit assignment problem.
- Temporal Difference (TD) learning is driven by the error/difference between temporally successive predictions; learning occurs whenever there is a change in prediction over time. [3]
- Temporal difference methods bootstrap from previous experience.

1.2.5.1 Q-Learning: Off-Policy TD-Learning

- Q-Learning is an off-policy Temporal Difference Learning method. It is also model-free. [4, 5].
- Off-policy refers to the fact that the value of the optimal policy is learnt independently of the agent's actions, as opposed to on-policy learning where the value of the policy being carried out by the agent is learnt [6].

1.2.5.2 SARSA: On-Policy TD-Learning

SARSA is an on-policy TD-Learning method.

1.3 Planning

- Planning involves reasoning on a model of an environment in order to produce a sequence of actions that will achieve a goal.[7].

1.3.1 A* Search

- A* Search [8] is not necessarily a Planning algorithm, rather it is a search algorithm. However, a simple Planning agent with the purpose of navigation in a discretised state space can use A* for planning.
- Heuristic
- Admissable

1.4 Exploration in Reinforcement Learning

- [9] distinguished exploration methods into two categories: directed and undirected. A more recent work [10] distinguished between reward-free and reward-based exploration.

1.4.1 Random

1.4.2 Optimism

1.4.3 Intrinsically Motivated

1.4.4 Deliberate

1.5 Related Work (Planning and Learning)

- DARLING
- Dyna
- AlphaGo?

Chapter 2

Methods

<Everything that comes under the 'Methods' criterion in the mark scheme should be described in one, or possibly more than one, chapter(s).>

2.1 Notation and Formalisms

2.2 Method Definition

2.3 Proof of Convergence

Chapter 3

Results

<Results, evaluation (including user evaluation) *etc.* should be described in one or more chapters. See the 'Results and Discussion' criterion in the mark scheme for the sorts of material that may be included here.>

3.1 A section

3.1.1 A sub-section

3.2 Another section

Chapter 4

Discussion

<Everything that comes under the 'Results and Discussion' criterion in the mark scheme that has not been addressed in an earlier chapter should be included in this final chapter. The following section headings are suggestions only.>

4.1 Conclusions

4.2 Ideas for future work

References

[1] M. Leonetti, L. Iocchi, and P. Stone, "A synthesis of automated planning and reinforcement learning for efficient, robust decision-making," *Artificial Intelligence*, vol. 241, pp. 103 – 130, September 2016.

[2] R. S. Sutton and A. G. Barto, *Reinforcement learning - an introduction*. Adaptive computation and machine learning, MIT Press, 1998.

[3] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, pp. 9–44, August 1988.

[4] C. J. C. H. Watkins, *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK, May 1989.

[5] C. J. C. H. Watkins and P. Dayan, "Technical note q-learning.," *Mach. Learn.*, vol. 8, pp. 279–292, 1992.

[6] D. Poole and A. Mackworth, *Artificial Intelligence: Foundations of Computational Agents*. Cambridge, UK: Cambridge University Press, 2 ed., 2017.

[7] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall, December 2002.

[8] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[9] S. Thrun, "Efficient exploration in reinforcement learning.," Tech. Rep. CMU-CS-92-102, Carnegie Mellon University, Pittsburgh, PA, January 1992.

[10] S. Amin, M. Gomrokchi, H. Satija, H. van Hoof, and D. Precup, "A survey of exploration methods in reinforcement learning," *CoRR*, vol. abs/2109.00157, 2021.

Appendix A

Self-appraisal

<This appendix should contain everything covered by the 'self-appraisal' criterion in the mark scheme. Although there is no length limit for this section, 2—4 pages will normally be sufficient. The format of this section is not prescribed, but you may like to organise your discussion into the following sections and subsections.>

A.1 Critical self-evaluation

A.2 Personal reflection and lessons learned

A.3 Legal, social, ethical and professional issues

<Refer to each of these issues in turn. If one or more is not relevant to your project, you should still explain *why* you think it was not relevant.>

A.3.1 Legal issues

A.3.2 Social issues

A.3.3 Ethical issues

A.3.4 Professional issues

Appendix B

External Material

<This appendix should provide a brief record of materials used in the solution that are not the student's own work. Such materials might be pieces of codes made available from a research group/company or from the internet, datasets prepared by external users or any preliminary materials/drafts/notes provided by a supervisor. It should be clear what was used as ready-made components and what was developed as part of the project. This appendix should be included even if no external materials were used, in which case a statement to that effect is all that is required.>