
	<p style="text-align: center;">THE UNIVERSITY OF SHEFFIELD Department of Electronic and Electrical Engineering MSc Individual Project Project Initialisation Document</p>			
Student Name	Yuche Huang			
Project Title	Reinforcement learning and control			
Supervisor	Dr. Peter Rockett	Second Marker	Dr. Mark Hopkinson	

Description and aims of Project:

In traditional control system design, developers intend to describe the behaviour in a system for a machine reacting as they expect. However, some the unknown situation or behaviours are difficult to pre-empt and designed due to diversity environment. Therefore, reinforcement learning (RL) that offers ability for the system to adapt its environment can be exploited to alleviate the problem. According the theory demonstrated by Darwin: species modify their action by interacting with the environment in order to survive and increase. Based on the theory, RL is a type of machine learning that allows agent to interact with its environment for solving the optimal and adaptive control by modifying its policies and action [1]. For instance, when we design a proportional integral derivative (PID) controller, some control parameters have to be set manually to find an appropriate value to minimize the responding time. By learning the action quality from the previous behaviour, RL modify the coefficients of PID control to optimize performance for the control system. The concept of the RL algorithm is that the system successful control decision should be remembered, and the idea can be defined by the Markov decision processes (MDP) in fig. 1.

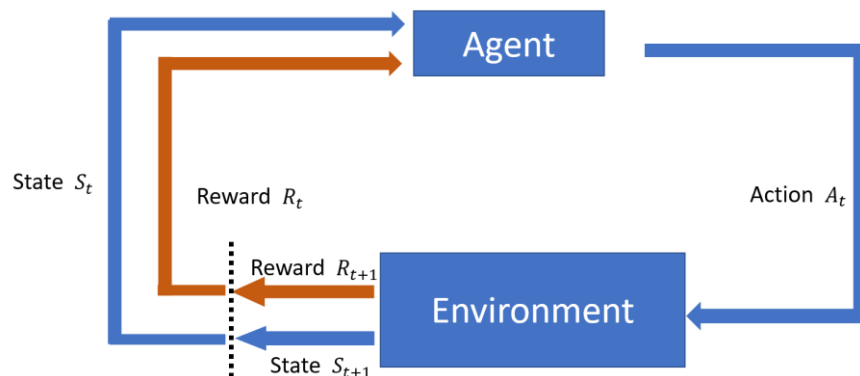


Fig. 1: The structure of Markov decision processes

In MDP model, the agent makes the corresponding action that depends on the state given from the environment. Simultaneously, the action that the agent has done is received by the environment and giving agent reward or punishment as a feedback signal to adjust the decision quality.

Although RL provides the ability for control system to adapt to the environment and optimize the performance, it still suffers from some problems. Due to RL algorithm is based on discrete-space of Bellman's equation, the algorithm can only operate in discrete state and action. However, some control system requires continuous state or variable such as speed, position, etc to operate more accurately. The purpose of the project is to research a method that provides algorithm operating in a continuous-state MDP for a control system.

Literature review:

The idea for reinforcement learning is from a fundamental role of animal learning that can interact with the environment and enhance its behaviour. According to the previous research [2], it demonstrates that RL can be distinguished in two common class, model-free and model-based, which operates in a different method for optimization. Model-based adopted the statistic method using previous experience to construct the model by receiving outcome reacting from the environment. Each information can be stored and analysed the relationship between reward and the action it has made statistically to generate an internal map for making an appropriate action next time. In addition, due to the model offers constant replanning, the agent can make an optimal decision even it is received contingencies event from the environment. Based on the characteristic, this model is suitable for goal-directed action that planned and purposeful [3].

In contrast, model-free RL using the previous experience directly to enhance the optimal operation without using neither the reward function nor transition function such as Q- function to estimate the reward value in the future. The condition to get the high reward occurs only if the outcomes with high utilities affected by the action which the agent takes immediately. Based on the characteristic, it can be exploited to alleviate an learning rule with momentary inconsistency as a error and offer the model to learn more accurate data and eliminate the inconsistencies. In addition, due to model-free uses the information directly rather than using the one that is combined with the previous information which is possibly erroneous about the state values, model-free has lower efficient than model-based [2].

In control community, it is extremely heard to using standard optimization methodology to solve the optimization unless we know the dynamics pattern. To alleviate this issue, using dynamic system such as Q-function plays a key role due to it offers a function to find a control policy recursively by beginning at the starting time and solving for polices at earlier times [4]. Q-function creates the relationship between actions and states. Every possible action is assigned a value by considering both the reward from taking the specific action and predicting the reward in the future based on the new state affected by the action agent has taken. This concept can be described in the following formula (1).

$$Q(x, u) = (1 - \alpha)Q(x, u) + \alpha(R + \gamma \max Q(x_{t+1}, u_{t+1})) \quad (1)$$

Based on the theory, rather than just finding the optimal policy, this method learns the values of all action. In addition, for Q-learning, any decision that agent has made can be executed at any time and information is improved form this experience. Therefore, Q-learning can learn some valuable data form other controllers even if the controllers are achieving a different goal of task.

To achieve that agent can vary the actions smoothly corresponding to smooth changes in state, the algorithm takes infinite number of states into account. The simplest methodology to solve a continuous-state is to discretize the state space. However, if we tend to discretize the state, the function will be representation by several block of constant value. In addition, according to the article [4], when discretizing each of n dimensions of states into k values, the total number of discrete state increases in exponential. Therefore, it leads to the states is far too many to represent.

Project Specification:

This project requires background knowledge of reinforcement learning which is a technique of machine learning and the theory of control systems such as PID control and Linear quadratic regulator. Firstly, reader should understand the principle of Markov decision processes which offers a function for program to adapt to the environment. The next step is learning dynamic programming such as Q-function which is a key factor to affect the decision policies and optimize the maximize reward. The last stage is to compare the learning performance with a different methodology such as using Model-based or model-free framework.

Project Schedule:

A – Background research on RL

B – Project initialisation document

C – Study the principle of RL

D – Background research on control theory

E – Second marker meeting

F – Algorithm Design (using Python)

G – Evaluation and performance improvement

H – Interim Report

I – Second viva

J – IEEE style Report

Gantt chart (use as a bar chart in conjunction with the main headings above)

Component	1	2	3	4	5	6	7	8	9	10	11	12	Exam Week	16	17	18	19	20	21	22
A																				
B																				
C																				
D																				
E																				
F																				
G																				
H																				
I																				
J																				

References:

Web references should be kept to a minimum as they are usually not peer-reviewed.

- [1] F. L. Lewis, "Learning and Adaptive Dynamic Programming for Feedback Control," 2009.
- [2] P. Dayan and Y. Niv, "Reinforcement learning : The Good , The Bad and The Ugly," pp. 1–12, 2008.
- [3] C. M. Gillan, A. R. Otto, E. A. Phelps, and N. D. Daw, "Model-based learning protects against

forming habits,” *Cogn. Affect. Behav. Neurosci.*, vol. 15, no. 3, pp. 523–536, 2015.

- [4] B. Recht, “A Tour of Reinforcement Learning: The View from Continuous Control,” pp. 1–28, 2018.

Risk Register:

Identify the key problems that could prevent your project from completing on time and associate a likeliness and risk level (Low/Medium/High). How can these risks be reduced?

Risk Number	Description of Risk	Mitigation of Risk	Risk evaluation (L/M/H)	Chance of risk occurring (L/M/H)
1	Loss of data (USB key)	Multiple back-ups in multiple locations	M	L
2	Laptop is broken	Back-up the file to google driver	M	L