Reinforcement Learning: An Introduction
Attempted Solutions
Chapter 3

Scott Brownlie & Rafael Rui

1 Exercise 3.1

Devise three example tasks of your own that fit into the MDP framework, identifying for each its states, actions, and rewards. Make the three examples as *different* from each other as possible. The framework is abstract and flexible and can be applied in many different ways. Stretch its limits in some way in at least one of your examples.

An e-commerce sit could use reinforcement learning to control daily pricing of products. The actions would be the the prices set for each product on each day. The states might include the month of the year, the day of the week and the proximity to special days such as Christmas and Valentine's Day. The reward would be the profit at the end of each day.

The manager of a football team could use reinforcement learning to pick the 11 players to play each game. The actions would be the team selection. The states could be the opponent, whether the game is home or away and the fitness of the players. The reward would be 0, 1 or 3 depending on whether the team lost, drew or won the game.

A company could use reinforcement learning to control the air temperature in its office. The actions would be the specific settings of the heating/air-conditioning system. The states would include the current outdoor temperature and the indoor temperature in each room in the building. The reward would be the satisfaction of the employees, which could be measured by selecting 10 employees at random every hour and asking them to rate the their comfort on a scale of 1 to 10 and then averaging the ratings.

2 Exercise 3.2

Is the MDP framework adequate to usefully represent all goal-directed learning tasks? Can you think of any clear exceptions?

Consider the example of a trader using reinforcement learning to buy and sell shares at the beginning of each day with the goal of maximising profit. A good trading strategy will take into account the prices of the shares over the last several days, perhaps even the last several months. Is this an example of a goal-directed learning task which the MDP framework does not usefully represent? Or can we include the historic share prices in the current state vector?