# CSE 518 - Artificial Intelligence
# Homework

Instructor: Shashi Prabh

## Chapter 22. Reinforcement Learning

**22.1**

- Implement a passive learning agent in a simple environment, such as the $4 \times 3$ world. For the case of an initially unknown environment model, compare the learning performance of the direct utility estimation and TD algorithms. Do the comparison for the optimal policy and for several random policies. For which do the utility estimates converge faster? What happens when the size of the environment is increased? (Try environments with and without obstacles.)
- Compute the expected utility of each state in the $4 \times 3$ world using the Value Iteration algorithm. Assume a negative reward of $-0.04$ for each step and a discount factor of 0.9.
- Implement the Q-learning algorithm to calculate the Q-values for each state-action pair in the $4 \times 3$ world. Experiment with different learning rates.
- Derive Q-values from the utilities computed in the second part of this exercise and compare them to those computed using Q-learning.

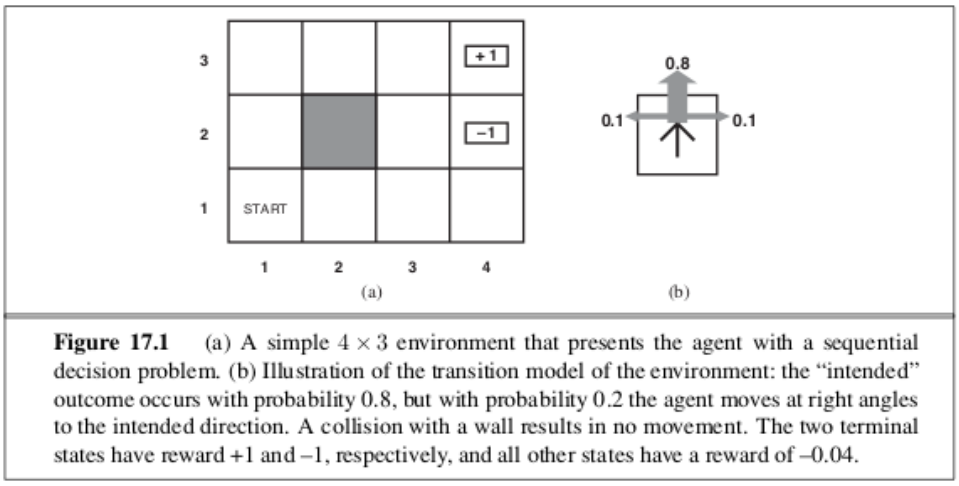


**Figure 17.1** (a) A simple $4 \times 3$ environment that presents the agent with a sequential decision problem. (b) Illustration of the transition model of the environment: the "intended" outcome occurs with probability 0.8, but with probability 0.2 the agent moves at right angles to the intended direction. A collision with a wall results in no movement. The two terminal states have reward +1 and –1, respectively, and all other states have a reward of –0.04.

Figure 1: Exercise 22.1

**22.2** Adapt the vacuum world for reinforcement learning by including rewards for squares being clean. Make the world observable by providing suitable percepts. Now experiment with different reinforcement

learning agents. Is function approximation necessary for success? What sort of approximator works for this application?

**22.3** Investigate the application of reinforcement learning ideas to the modeling of human and animal behavior.

**22.4** Is reinforcement learning an appropriate abstract model for evolution? What connection exists, if any, between hardwired reward signals and evolutionary fitness?