CS255: Artificial Intelligence
Seminar Sheet 8 — Reinforcement Learning[1]

1. Suppose S is a set of belief states, P a set of possible percepts, and C a set of possible commands, such that $c_t = command(s_t, p_t)$ means that the agent controller issues command c_t when the belief state is s_t and p_t is observed. Suppose the agent uses Q-learning with discount factor γ, step size of α and is carrying out an ϵ-greedy exploration strategy.

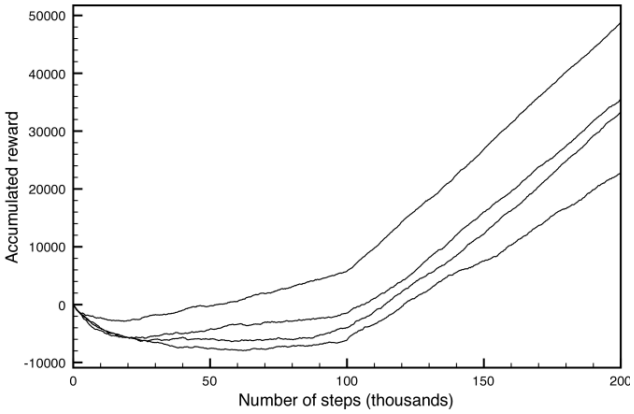
   (a) What are the components of the belief state of the Q-learning agent?

   (b) What are the percepts?

   (c) What is the command function of the Q-learning agent?

   (d) What is the belief-state transition function of the Q-learning agent?

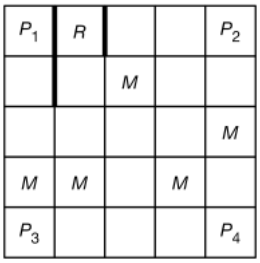
2. For the plot of the total reward as a function of time given below, the minimum and zero crossing are only meaningful statistics when balancing positive and negative rewards is reasonable behaviour. Suggest what should replace these statistics when zero reward is not an appropriate definition of reasonable behaviour.



3. Explain what happens in reinforcement learning if the agent always chooses the action that maximizes the Q-value. Suggest two ways to force the agent to explore.

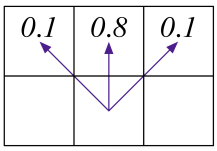
4. Consider the grid game from Example 12.2 in AIFCA, illustrated below, in which the P_i's represent prizes which may be present, M represents monsters that can appear at any time (and damage the agent if they occur the same location), and R is a repair station for the agent to fix itself. Suppose the agent has access to features corresponding to the x and y distances to the current prizes. Since these features do not depend on the action, are they of use for Q-learning?

[1] Exercises taken from Artificial Intelligence: Foundations of Computational Agents, Poole and Mackworth, 2017, with minor adaptations.

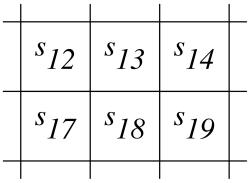


5. Suppose a Q-learning agent, with fixed α and discount γ, was in state 34, did action 7, received reward 3, and ended up in state 65. What value(s) get updated? Give an expression for the new value(s).

6. Consider a grid world where the action "up" has the following dynamics:



That is, it goes up with probability 0.8, up-left with probability 0.1, and up-right with probability 0.1. Suppose we have the following states:



There is a reward of $+10$ upon entering state $s14$, and a reward of -4 upon entering state $s15$. All other rewards are 0.

Suppose you are doing Q-learning with learning rate $\alpha = 0.1$ and discount factor $\gamma = 0.95$. Initially all Q-values are zero.

(a) Suppose initially you have following sequence of state-action-reward experiences:

$$s_{17}, right, 0, s_{18}, up, 10, s_{14}, right, -4, s_{15}$$

Show what Q values are assigned due to this sequence of experiences.

(b) Suppose that later you have the following sequence of state-action-reward experiences (and that you had not visited these states in between these times):

$$s_{23}, up, 0, s_{18}, up, 0, s_{13}, right, 10, s_{14}$$

Show what Q values are assigned due to this sequence of experiences.