COMPUTER SCIENCE TRIPOS Part IA – 2012 – Paper 7

2 Artificial Intelligence II (SBH)

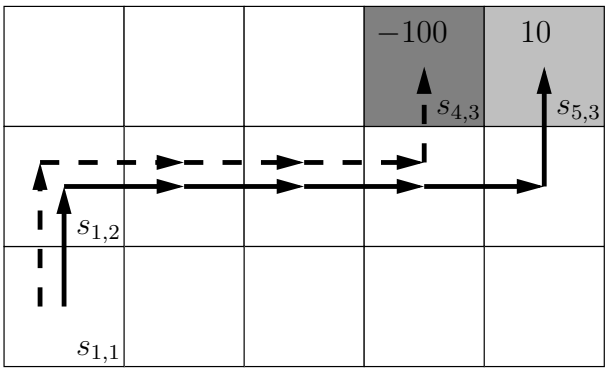
A *reinforcement learning problem* has states $\{s_1, \ldots, s_n\}$, actions $\{a_1, \ldots, a_m\}$, reward function $R(s, a)$ and next state function $S(s, a)$.

(a) Give a general definition of a *policy* for such a problem. [1 mark]

(b) Give a general definition of the *discounted cumulative reward* and the corresponding *optimal policy* for such a problem. [5 marks]

(c) Give an expression for the optimal policy in terms of R, S and the discounted cumulative reward, and show how this can be modified to produce the *Q-learning algorithm*. [7 marks]

In a simple reinforcement learning problem, states are positions on a grid and actions are `up` and `right`. The only way an agent can receive a reward is by moving into one of two special positions, one of which has a reward of 10 and the other of -100.



Here, states are labelled by their grid coordinates. A possible sequence of actions (sequence 1) is shown by solid arrows, ending with a reward of 10 being received, and another (sequence 2) by dashed arrows ending with a reward of -100.

(d) Assume that all Q values are initialised at 0.

(i) Explain how the Q values are altered if sequence 1 is used *twice* in succession by the Q-learning algorithm. [4 marks]

(ii) Explain what further changes occur to the Q values if sequence 2 is then used *once* by the Q-learning algorithm. [3 marks]