2 Artificial Intelligence II (SBH)

Consider a reinforcement learning problem having 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 discounted cumulative reward, a policy, and an optimal policy for a problem of this kind. [5 marks]

(b) Give a detailed derivation of the \(Q\)-learning algorithm. [5 marks]

(c) In the reinforcement learning problem shown in the diagram, states are positions on a grid and actions are down and right. The initial state is \(s_1\). The only way an agent can receive a (non-zero) reward is by moving into one of two special positions, one of which has reward \(-50\) and the other \(150\).

A possible sequence of actions (sequence 1) is shown by solid arrows, and another (sequence 2) by dashed arrows. Assume that all \(Q\) values are initialised at 0. Explain how the \(Q\) values are modified by the \(Q\)-learning algorithm if sequence 1 is used once, followed by two uses of sequence 2, and then one final use of sequence 1. [10 marks]