COMP9444 Neural Networks and Deep Learning
Term 3, 2019

Solutions to Exercise 6: Reinforcement Learning

This page was last updated: 11/06/2019 22:09:02

Consider an environment with two states $S = \{S_1, S_2\}$ and two actions $A = \{a_1, a_2\}$, where the (deterministic) transitions δ and reward R for each state and action are as follows:

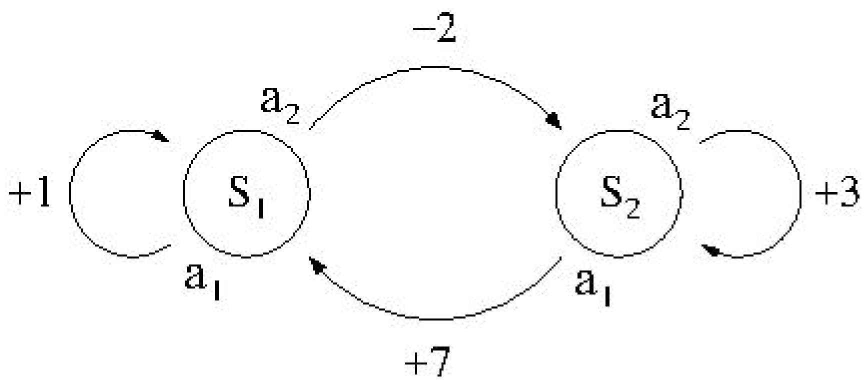
$\delta(S_1, a_1) = S_1,\ \ R(S_1, a_1) = +1$

$\delta(S_1, a_2) = S_2,\ \ R(S_1, a_2) = -2$

$\delta(S_2, a_1) = S_1,\ \ R(S_2, a_1) = +7$

$\delta(S_2, a_2) = S_2,\ \ R(S_2, a_2) = +3$

1. Draw a picture of this environment, using circles for the states and arrows for the transitions.



2. Assuming a discount factor of $\gamma = 0.7$, determine:
   a. the optimal policy $\pi^* : S \to A$

      $\pi^*(S_1) = a_2$

      $\pi^*(S_2) = a_1$

   b. the value function $V : S \to R$

      $V(S_1) = -2 + \gamma V(S_2)$

      $V(S_2) = +7 + \gamma V(S_1)$

      So $\quad V(S_1) = -2 + 7\gamma + \gamma^2 V(S_1)$

      i.e. $\quad V(S_1) = (-2 + 7\gamma)/(1 - \gamma^2) = (-2 + 7 \times 0.7)/(1 - 0.49) = 5.69$

      Then $\quad V(S_2) = 7 + 0.7 \times 5.69 = 10.98$

   c. the "Q" function $Q : S \times A \to R$

$Q(S_1, a_1) = 1 + \gamma V(S_1) = 4.98$

$Q(S_1, a_2) = V(S_1) = 5.69$

$Q(S_2, a_1) = V(S_2) = 10.98$

$Q(S_2, a_2) = 3 + \gamma V(S_2) = 10.69$

Writing the Q values in a matrix, we have:

Q	a_1	a_2
S_1	4.98	5.69
S_2	10.98	10.69

Trace through the first few steps of the Q–learning algorithm, with a learning rate of 1 and with all Q values initially set to zero. Explain why it is necessary to force exploration through probabilistic choice of actions, in order to ensure convergence to the true Q values.

With a deterministic environment and a learning rate of 1, the Q–Learning update rule is

$Q(S, a) \leftarrow r(S, a) + \gamma \max_b Q(\delta(S, a), b)$

Let's assume the agent starts in state S_1. Since the initial Q values are all zero, the first action must be chosen randomly. If action a_1 is chosen, the agent will get a reward of +1 and update
$Q(S_1,a_1) \leftarrow 1 + \gamma \times 0 = 1$

If we do not force exploration, the agent will always prefer action a_1 in state S_1, and will never explore action a_2. This means that $Q(S_1, a_2)$ will remain zero forever, instead of converging to the true value of 5.69 . If we do force exploration, the next steps may look like this:

current state	chosen action	new Q value
S_1	a_2	$-2 + \gamma*0 = -2$
S_2	a_2	$+3 + \gamma*0 = +3$

At this point, the table looks like this:

Q	a_1	a_2
S_1	1	−2
S_2	0	3

Again, we need to force exploration, in order to get the agent to choose a_1 from S_2, and to again choose a_2 from S_1

current state	chosen action	new Q value
S_2	a_1	$+7 + \gamma*1 = 7.7$
S_1	a_2	$-2 + \gamma*7.7 = 3.39$

Q	a_1	a_2
S_1	1	3.39
S_2	7.7	3

Further steps will refine the Q value estimates, and, in the limit, they will converge to their true values.

3. Now let's consider how the Value function changes as the discount factor γ varies between 0 and 1.
   There are four deterministic policies for this environment, which can be written as π_{11}, π_{12}, π_{21} and π_{22}, where $\pi_{ij}(S_1) = a_i$, $\pi_{ij}(S_2) = a_j$

   a. Calculate the value function $V^{\pi}{}_{(\gamma)}: S \to R$ for each of these four policies (keeping γ as a variable)

      $V^{\pi}{}_{11}(S_1) = +1 + \gamma\, V^{\pi}{}_{11}(S_1)$, so $V^{\pi}{}_{11}(S_1) = 1/(1 - \gamma)$
      $V^{\pi}{}_{11}(S_2) = +7 + \gamma\, V^{\pi}{}_{11}(S_1) = 7 + \gamma/(1 - \gamma)$

      $V^{\pi}{}_{12}(S_1) = V^{\pi}{}_{11}(S_1) = 1/(1 - \gamma)$
      $V^{\pi}{}_{12}(S_2) = 3/(1 - \gamma)$

      $V^{\pi}{}_{21}(S_1) = -2 + 7\gamma + \gamma^2 V^{\pi}{}_{21}(S_1)$, so $V^{\pi}{}_{21}(S_1) = (-2 + 7\gamma)/(1 - \gamma^2)$
      $V^{\pi}{}_{21}(S_2) = +7 - 2\gamma + \gamma^2 V^{\pi}{}_{21}(S_2)$, so $V^{\pi}{}_{21}(S_2) = (7 - 2\gamma)/(1 - \gamma^2)$

      $V^{\pi}{}_{22}(S_1) = -2 + 3\gamma/(1 - \gamma)$
      $V^{\pi}{}_{22}(S_2) = 3/(1 - \gamma)$

   b. Determine for which range of values of γ each of the policies π_{11}, π_{12}, π_{21}, π_{22} is optimal

      π_{11} is optimal when
      $0 < V^{\pi}{}_{11}(S_1) - V^{\pi}{}_{21}(S_1) = ((1 + \gamma) - (-2 + 7\gamma))/(1 - \gamma^2) = (3 - 6\gamma)/(1 - \gamma^2)$,
      i.e. $0 \leqslant \gamma \leqslant 0.5$

      π_{22} is optimal when
      $0 < V^{\pi}{}_{22}(S_2) - V^{\pi}{}_{21}(S_2) = (3(1 + \gamma) - (7 - 2\gamma))/(1 - \gamma^2) = (-4 + 5\gamma)/(1 - \gamma^2)$, i.e. $0.8 \leqslant \gamma < 1.0$

      π_{21} is optimal for $0.5 \leqslant \gamma \leqslant 0.8$

π_{12} is never optimal because it is dominated by π_{11} when $\gamma < 2/3$ and by π_{22} when $\gamma > 0.6$
