Assignment 4: Reinforcement Learning

Sébastien Bouquet

1 Task 1 - Basic probabilities

a)

The relative probability of being caught $P(C)$ is the sum of the probabilities to be caught in any given state times the probability of being in that state : $P(C) = \sum_x P(C|S=x) = 0.4 * 0.1 + 0.3 * 0.4 + 0.2 * 0.8 + 0.1 * 0.2 = \mathbf{0.34}$.

b)

The proportion of lizard that are fighting of those that were caught is equal to the probability of a lizard to be fighting given that it was caught :
$P(S=F|C) = \frac{P(S=F \cap C}{P(C)} = \frac{0.2*0.8}{0.34} \approx \mathbf{0.47} = \mathbf{47}\%$.

2 Task 2 - Markov Decision Process

1.

a) Yes, there is the a problem with the reward function: since the robots receives +1 reward each time step it stays in the maze and it tries to maximize its total reward, its optimal strategies is to stay indefinitely in the maze since it can expect an infinite reward this way and $\infty > 1000$ (the reward to reach the end being 1000.)

b) It can be solved by not giving the robot a reward for simply staying in the maze. Instead it should receive a reward only when it makes a move that gets it closer to the end of the maze. For example, we could define that if it is one step from the exit, it receives +1000 reward for reaching the exit and -1 reward for going back or turning left or right; then, if it is two steps from the exit, it receives +999 reward for reaching the state that is one step from the exit, etc.
In effect, it receives a negative (eg. -1) reward for staying in the maze instead of a +1 reward.

4.

a)

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... \\ &= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + ...) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned} \tag{1}$$

b) The sufficient condition for the series to converge is $\gamma < 1$ and R_t non zero and constant for all t.

c) $G_t = \sum_{k=0}^{\infty} \gamma^k * R_{t+k+1}$
since $R_t = 1 \; \forall \; t$ then $\sum_{k=0}^{\infty} 1 * \gamma^k = \frac{1}{1-\gamma}$ for $|\gamma| < 1$

d) $G_5 = R_5 = 0$
$G_4 = R_5 + \gamma G_5 = 6 + 0.3 * 0 = 6$
$G_3 = R_4 + G_4 = 1 + 0.3 * 6 = 2.8$
$G_2 = R_3 + G_3 = 2 + 0.3 * 2.8 = 2.84$
$G_1 = R_2 + G_2 = -0.5 + 0.3 * 2.84 = 0.352$
$G_0 = R_1 + G_1 = -1 + 0.3 * 0.352 = -0.8944$

e) $G_t = \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) = \sum_{k=0}^{\infty} \gamma^k * c + \sum_{k=0}^{\infty} \gamma^k * r = \frac{c}{1-\gamma} + \frac{r}{1-\gamma} = \frac{c+r}{1-\gamma}$
(considering R_t non zero, constant and equal to r for all t)

f) $G_t = \sum_{k=0}^{T-1} \gamma^k (R_{t+k+1} + c) = \sum_{k=0}^{T-1} c * \gamma^k + \sum_{k=0}^{T-1} \gamma^k * r = c\Big(\frac{1-\gamma^T}{1-\gamma}\Big) + r\Big(\frac{1-\gamma^T}{1-\gamma}\Big) = \frac{(c+r)(1-\gamma)}{1-\gamma} = c + r$, considering R_t non zero, constant and equal to r for all t.

3 Task 3 - Dynamic programming

1. $v_*(s) = \max\limits_a \sum\limits_{s',r} p(s', r \mid s, a) * \big(r + \gamma v_*(s')\big)$

2. $q_*(s, a) = \sum\limits_{s',r} p(s', r \mid s, a) * \big(r + \gamma \max\limits_{a'} q_*(s', a')\big)$

3. (a) **Iteration 1**
Policy iteration
Init: $V(s) = 0$ for all $s \in S$, $\gamma = 1$, $\pi_0(s) = (U, D, R, D, L, L, U, D, L, R, U, D, R, L)$, $r = -1$.
Evaluation:
For each $s \in S$:
$V(s) = \sum\limits_{s',r} p(s', r|s, \pi(s)) * (r + \gamma V(s'))$
$V(1) = p(s' = 1, r|1, \pi_0(1)) * (r + V(1)) = 1 * (-1 + 0) = -1$ (every state s' st $s' \neq 1$ is not in the summation because $p(s' \neq 1, r|1, \pi_0(1))) = 0$

$V(2) = p(s' = 3, r|2, \pi_0(2)) * (r + V(3)) = 1 * (-1 + 0) = -1$

$V(3) = p(s' = 3, r|3, \pi_0(3)) * (r + V(3)) = 1 * (-1 + 0)) = -1$

$V(4) = p(s' = 8, r|4, \pi_0(4)) * (r + V(8)) = 1 * (-1 + 0) = -1$

$V(5) = p(s' = 4, r|5, \pi_0(5)) * (r + V(4)) = 1 * (-1 + -1) = -2$

$V(6) = p(s' = 5, r|6, \pi_0(6)) * (r + V(5)) = 1 * (-1 + -1) = -3$

$V(7) = p(s' = 3, r|7, \pi_0(7)) * (r + V(3)) = 1 * (-1 + -1) = -2$

$V(8) = p(s' = 12, r|8, \pi_0(12)) * (r + V(12)) = 1 * (-1 + 0) = -1$

$V(9) = p(s' = 8, r|9, \pi_0(9)) * (r + V(8)) = 1 * (-1 + -1) = -2$

$V(10) = p(s' = 11, r|10, \pi_0(10)) * (r + V(11)) = 1 * (-1 + 0) = -1$

$V(11) = p(s' = 7, r|11, \pi_0(11)) * (r + V(7)) = 1 * (-1 + -2) = -3$

$V(12) = p(s' = 12, r|12, \pi_0(12)) * (r + V(12)) = 1 * (-1 + 0) = -1$

$V(13) = p(s' = 14, r|13, \pi_0(13)) * (r + V(14)) = 1 * (-1 + 0) = -1$

$V(14) = p(s' = 13, r|14, \pi_0(14)) * (r + V(13)) = 1 * (-1 + -1) = -2$

$\Rightarrow V(s) = (0, -1, -1, -1, -1, -2, -3, -2, -1, -2, -1, -3, -1, -1, -2, 0)$

($V(0)$ and $V(16)$ are fixed to 0)

Improvement:

$\pi_1(s) = argmax_a \sum_{s',r} p(s', r|s, a)(r + V(s')$

$\pi_1(s) = (L, R, R, U, L, U, U, D, L, D, D, D, L, R)$

Iteration 2

Evaluation:

$V(s) = (0, -1, -2, -2, -1, -2, -3, -3, -2, -3, -3, -1, -2, -3, -1, 0)$

Improvement:

$\pi_2(s) = (L, L, R, U, L, U, D, U, L, R, D, D, R, R)$

Iteration 3

Evaluation:

$V(s) = (-1, -2, -3, -1, -2, -3, -2, -2, -3, -2, -1, -3, -2, -1)$

Improvement:

$\pi_3(s) = (L, L, D, U, L, U, D, U, L, R, D, R, R, R)$

(b) **Iteration 1**

Init: $V(s) = 0$ for all $s \in S$, $\gamma = 1$, $r = -1$.

$V(s) = max_a \sum_{s}^{\prime}, rp(s', r|s, a) * (r + V(s'))$

(here again only one term of the summation is important since $p(s', r|s, a) = 1$ for one $a, 0$ for the other.)

$V(1) = max_a p(s' = a, r|1, a) * (r + V(s')) = 1 * (-1 + 0) = -1, s' \in \{0, 1, 2, 5\}$

(since $V(s) = 0 \,\forall\, s$, $V(1) = -1 \,\forall\, a$)

$V(2) = max_a p(s' = a, r|2, a) * (r + V(s')) = 1 * (-1 + 0) = -1, s' \in \{2, 3, 6\}$

$V(3) = max_a p(s' = a, r|3, a) * (r + V(s')) = 1 * (-1 + 0)) = -1, s' \in \{3, 7\}$

$V(4) = max_a p(s' = a, r|4, a) * (r + V(s')) = 1 * (-1 + 0) = -1, s' \in \{0, 4, 5, 8\}$

.

.

.

$V(14) = max_a p(s' = a, r|14, a) * (r + V(s')) = 1 * (-1 + 0) = -1, s' \in \{14, 15\}$

$\Rightarrow V(s) = -1 \; \forall \, s$

Iteration 2

$V(1) = max_a p(s' = a, r|1, a) * (r + V(s')) = 1 * (-1 + 0) = -1, s' = 0$

$V(2) = max_a p(s' = a, r|2, a) * (r + V(s')) = 1 * (-1 + -1) = -2, s' \in \{1, 2, 3, 6\}$

$V(3) = max_a p(s' = a, r|3, a) * (r + V(s')) = 1 * (-1 + -1)) = -2, s' \in \{3, 7\}$

$V(4) = max_a p(s' = a, r|4, a) * (r + V(s')) = 1 * (-1 + 0) = -1, s' = 0$

$V(5) = max_a p(s' = a, r|1, a) * (r + V(s')) = 1 * (-1 + -1) = -2, s' \in \{1, 4, 6, 9\}$

$V(6) = max_a p(s' = a, r|2, a) * (r + V(s')) = 1 * (-1 + -1) = -2, s' \in \{7, 10\}$

$V(7) = max_a p(s' = a, r|3, a) * (r + V(s')) = 1 * (-1 + -1)) = -2, s' \in \{7, 11\}$

$V(8) = max_a p(s' = a, r|4, a) * (r + V(s')) = 1 * (-1 + -1) = -2, s' \in \{8, 9, 12\}$

$V(9) = max_a p(s' = a, r|3, a) * (r + V(s')) = 1 * (-1 + -1)) = -2, s' \in \{10, 13\}$

$V(10) = max_a p(s' = a, r|4, a) * (r + V(s')) = 1 * (-1 + -1) = -2, s' \in \{11, 14\}$

$V(11) = max_a p(s' = a, r|1, a) * (r + V(s')) = 1 * (-1 + 0) = -1, s' = 15$

$V(12) = max_a p(s' = a, r|2, a) * (r + V(s')) = 1 * (-1 + -1) = -2, s' \in \{12, 13\}$

$V(13) = max_a p(s' = a, r|3, a) * (r + V(s')) = 1 * (-1 + -1)) = -2, s' \in \{13, 14\}$

$V(14) = max_a p(s' = a, r|4, a) * (r + V(s')) = 1 * (-1 + 0) = -1, s' = 15$

Iteration 3

$V(1) = max_a p(s' = a, r|1, a) * (r + V(s')) = 1 * (-1 + 0) = -1, s' = 0$

$V(2) = max_a p(s' = a, r|2, a) * (r + V(s')) = 1 * (-1 + -1) = -2, s' = 1$

$V(3) = max_a p(s' = a, r|3, a) * (r + V(s')) = 1 * (-1 + -1)) = -3, s' \in \{2, 3, 7\}$

$V(4) = max_a p(s' = a, r|4, a) * (r + V(s')) = 1 * (-1 + 0) = -1, s' = 0$

$V(5) = max_a p(s' = a, r|1, a) * (r + V(s')) = 1 * (-1 + -1) = -2, s' \in \{1, 4\}$

$V(6) = max_a p(s' = a, r|2, a) * (r + V(s')) = 1 * (-1 + -1) = -3, s' \in \{2, 5, 7, 10\}$

$V(7) = max_a p(s' = a, r|3, a) * (r + V(s')) = 1 * (-1 + -1)) = -2, s' = 11$

$V(8) = max_a p(s' = a, r|4, a) * (r + V(s')) = 1 * (-1 + -1) = -2, s' = 4$

$V(9) = max_a p(s' = a, r|3, a) * (r + V(s')) = 1 * (-1 + -1)) = -3, s' \in \{5, 8, 10, 13\}$

$V(10) = max_a p(s' = a, r|4, a) * (r + V(s')) = 1 * (-1 + -1) = -2, s' \in \{11, 14\}$

$V(11) = max_a p(s' = a, r|1, a) * (r + V(s')) = 1 * (-1 + 0) = -1, s' = 15$

$V(12) = max_a p(s' = a, r|2, a) * (r + V(s')) = 1 * (-1 + -1) = -3, s' \in \{8, 12, 13\}$

$V(13) = max_a p(s' = a, r|3, a) * (r + V(s')) = 1 * (-1 + -1)) = -2, s' = 14$

$V(14) = max_a p(s' = a, r|4, a) * (r + V(s')) = 1 * (-1 + 0) = -1, s' = 15$