calculations

Lior Fuks and Markus Merklinger

January 2018

1 Calculation Exercises

The formula for updating the Q matrix is

$$Q(s_t, a_t) \leftarrow (1-\alpha) \cdot \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \overbrace{\left(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} \right)}^{\text{learned value}}$$

(source: Wikipedia)
we use learning rate of 1 and a discount factor of 0.5. We receive this Q update formula:

$$Q(s_t, a_t) \leftarrow \left(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} \right)$$

The given reward matrix is: $r(s) = \begin{pmatrix} -1 & -1 & 0 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \end{pmatrix}$

And the initial Q matrix is: $Q = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

Starting at S(1,1) at taking action: down:

$$Q(s_t, a_t) \leftarrow \Big(-1 + 0.5 \cdot 0 \Big) \qquad \text{leads to } Q = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

from s(2,1) action 2: right:

$$Q(s_t, a_t) \leftarrow \Big(-1 + 0.5 \cdot 0 \Big) \qquad \text{leads to } Q = \begin{pmatrix} -1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

from s(2,2) action 3: up:

$$Q(s_t, a_t) \leftarrow \Big(-1 + 0.5 \cdot 0 \Big) \qquad \text{leads to } Q = \begin{pmatrix} -1 & 0 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

from s(2,2) action 4: right:

$$Q(s_t, a_t) \leftarrow \Big(-1 + 0.5 \cdot 0 \Big) \qquad \text{leads to } Q = \begin{pmatrix} -1 & 0 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

from s(2,3) action 5: up:

$$Q(s_t, a_t) \leftarrow \Big(-1 + 0.5 \cdot 0 \Big) \qquad \text{leads to } Q = \begin{pmatrix} -1 & 0 & 0 \\ -1 & -1 & -1 \\ 0 & 0 & 0 \end{pmatrix}$$