Artificial Neural Networks (Gerstner). Solutions for week 7

Reinforcement Learning: Q-value and SARSA

Exercise 1. Iterative update

We consider an empirical evaluation of $Q(s,a)$ by averaging the rewards for action a over the first k trials:

$$Q_k = \frac{1}{k}\sum_{i=1}^{k} r_i .$$

We now include an additional trial and average over all $k+1$ trials.

a. Show that this procedure leads to an iterative update rule of the form

$$\Delta Q_k = \eta (r_k - Q_{k-1}),$$

(assuming $Q_0 = 0$).

b. What is the value of η?

c. Give an intuitive explanation of the update rule. *Hint: Think of the following: If the actual reward is larger than my estimate, then I should ...*

Solution:

a. We define ΔQ_k as the difference between Q_k and Q_{k-1}, and we simplify:

$$\begin{aligned}
\Delta Q_k = Q_k - Q_{k-1} &= \frac{1}{k}\sum_{i=1}^{k} r_i - \frac{1}{k-1}\sum_{i=1}^{k} r_i \\
&= \frac{1}{k}\left(r_k + \sum_{i=1}^{k-1} r_i\right) - \frac{1}{k-1}\sum_{i=1}^{k-1} r_i \\
&= \frac{1}{k}\left(r_k + \frac{k-1}{k-1}\sum_{i=1}^{k-1} r_i - \frac{k}{k-1}\sum_{i=1}^{k-1} r_i\right) \\
&= \frac{1}{k}\left(r_k - \frac{1}{k-1}\sum_{i=1}^{k-1} r_i\right) \\
&= \eta\left(r_k - Q_{k-1}\right) .
\end{aligned}$$

b. $\eta = 1/k$.

c. If the actual reward is larger than my estimate, then I should increase my estimate, otherwise I should decrease it.

Exercise 2. Greedy policy and the two-armed bandit

In the “2-armed bandit” problem, one has to choose one of 2 actions. Assume action a_1 yields a reward of $r = 1$ with probability $p = 0.25$ and 0 otherwise. If you take action a_2, you will receive a reward of $r = 0.4$ with probability $p = 0.75$ and 0 otherwise. The “2-armed bandit” game is played several times.

a. Assume that you initialize all Q values at zero. You first try both actions: in trial 1 you choose a_1 and get $r = 1$; in trial 2 you choose a_2 and get $r = 0.4$. Update your Q values ($\eta = 0.2$).

b. In trials 3 to 5, you play greedy and always choose the action which looks best (i.e., has the highest Q-value). Which action has the higher Q-value after trial 5?

c. Calculate the expected reward for both actions. Which one is the best?

d. Initialize both Q-values at 2 (optimistic). Assume that, as in the first part, in the first two trials you get for both actions the reward. Update your Q values once with $\eta = 0.2$. Suppose now that in the following rounds, in order to explore well, you choose actions a_1 and a_2 alternatingly and update the Q-values with a very small learning rate ($\eta = 0.001$). How many rounds (one round = two trials = one trial with each action) does it take *on average*, until the maximal Q-value also reflects the best action?
Hint: For $\eta \ll 1$ we can approximate the actual returns r_t with their expectations $E[r]$.

Solution:

a. In the beginning, $Q(a_1, t = 0) = Q(a_2, t = 0) = 0$ (we dropped the state index s since there is only a single state). After choosing action a_1 and receiving a reward of $r = 1$, its Q-value is updated to:

$$Q(a_1, t = 1) = Q(a_1, t = 0) + \Delta Q(a_1) = 0 + \eta(r - Q(a_1, t = 0)) = 0 + 0.2 \cdot 1 = 0.2.$$

After choosing action a_2 and receiving a reward of $r = 0.4$, its Q-value is updated to:

$$Q(a_2, t = 2) = Q(a_2, t = 0) + \Delta Q(a_2) = 0 + \eta(r - Q(a_2, t = 0)) = 0 + 0.2 \cdot 0.4 = 0.08.$$

Continuing with a greedy method implies that in the next round, action a_1 will be chosen.

b. In trial 3 you take action a_1. If the return is 0,

$$Q(a_1, t = 3) = Q(a_1, t = 2) + \eta(r - Q(a_1, t = 2)) = (1 - \eta) \cdot Q(a_1, t = 2) + \eta r = 0.8 \cdot 0.2 = 0.16\,.$$

Thus, in trial 4 we take again action a_1. If the return is again 0,

$$Q(a_1, t = 4) = (1 - \eta) \cdot Q(a_1, t = 3) = 0.8 \cdot 0.16 = 0.128\,.$$

In trial 5 we take again action a_1. If the return is again 0,

$$Q(a_1, t = 5) = (1 - \eta) \cdot Q(a_1, t = 4) = 0.8 \cdot 0.128 = 0.1024\,.$$

Thus, with a greedy policy, also in trial 6 action a_1 will be taken. If by chance some of the returns were 1 instead of 0, $Q(a_1, t = 5)$ would be even higher, while $Q(a_2, t = 5) = Q(a_2, t = 2) = 0.08$ because action a_2 was never taken.

c. For action a_1, the expected reward per round is given by $E[r_1] = p \cdot 1 + (1 - p) \cdot 0 = 0.25$. For action a_2, the expected reward per round is evaluated to $E[r_2] = 0.75 \cdot 0.4 + 0.25 \cdot 0 = 0.3$. The second action yields a higher reward on average.

d. Similarly as in a, we can compute the Q-values after the first step with $\eta = 0.2$. We obtain: $Q^*(a_1) = 1.8$ and $Q^*(a_2) = 1.68$. We use the hint that for $\eta \ll 1$ we can approximate the actual returns r_t with their expectations $E[r]$, i.e.

$$\begin{aligned} Q(a_i, t) &= (1 - \eta)Q(a_i, t - 1) + \eta r_t && (1)\\ &\approx (1 - \eta)Q(a_i, t - 1) + \eta E[r] && (2)\\ &= (1 - \eta)\left[(1 - \eta)Q(a_i, t - 2) + \eta E[r]\right] + \eta E[r] && (3) \end{aligned}$$

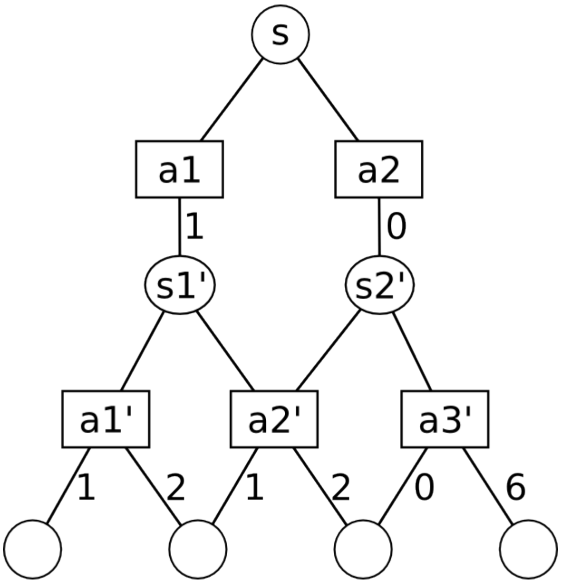

Figure 1: A tree–like environment.

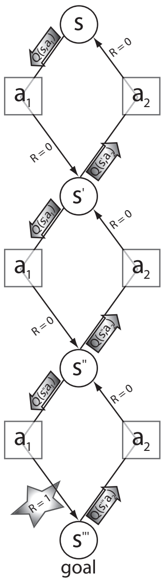

Figure 2: A linear maze.

We continue by induction and arrive at

$$
\begin{aligned}
(1-\eta)\left[(1-\eta)Q(a_i,t-2)+\eta E[r]\right]+\eta E[r] &= \eta\sum_{s=0}^{t-1}(1-\eta)^s E[r]+(1-\eta)^t Q(a_i,0) && (4)\\
&= \eta\frac{1-(1-\eta)^t}{\eta}E[r]+(1-\eta)^t Q(a_i,0) && (5)\\
&= (1-\gamma^t)E[r]+\gamma^t Q(a_i,0), && (6)
\end{aligned}
$$

with $\gamma = 1-\eta$ and using the formula for the geometric series. We search for the smallest t such that

$$
\begin{aligned}
Q(a_2,t) &> Q(a_1,t) && (7)\\
(1-\gamma^t)E[r_2]+\gamma^t Q^*(a_2) &> (1-\gamma^t)E[r_1]+\gamma^t Q^*(a_1) && (8)\\
\Rightarrow \gamma^t\left(Q^*(a_2)-Q^*(a_1)+E[r_1]-E[r_2]\right) &> E[r_2]-E[r_1] && (9)\\
\Rightarrow t\log(\gamma) &> \log\left(\frac{E[r_2]-E[r_1]}{Q^*(a_2)-Q^*(a_1)+E[r_1]-E[r_2]}\right) && (10)\\
\Rightarrow t &> 1223.16 && (11)
\end{aligned}
$$

Exercise 3. SARSA algorithm

In the lecture, we introduced the SARSA (state-action-reward-state-action) algorithm, which (for discount factor $\gamma = 1$) is defined by the update rule

$$\Delta Q(s,a) = \eta\left[r - \big(Q(s,a) - Q(s',a')\big)\right] , \qquad (12)$$

where s' and a' are the state and action subsequent to s and a. In this exercise, we apply a greedy policy, i.e., at each time step, the action chosen is the one with maximal expected reward, i.e.,

$$a_t^* = \arg\max_a Q_a(s,a) \, . \qquad (13)$$

If the available actions have the same Q-value, we take both actions with probability 0.5.

Consider a rat navigating in a 1-armed maze (=linear track). The rat is initially placed at the upper end of the maze (state s), with a food reward at the other end. This can be modeled as a one-dimensional sequence of states with a unique reward ($r = 1$) as the goal is reached. For each state, the possible actions are going up or going down (Fig. 2). When the goal is reached, the trial is over, and the rat is picked up by the experimentalist and placed back in the initial position s and the exploration starts again.

a. Suppose we discretize the linear track by 6 states, $s_1, \dots, s_6$. Initialize all the Q-values at zero. How do the Q-values develop as the rat walks down the maze in the first trial?

b. Calculate the Q-values after 3 complete trials. How many Q-value values are non-zero? How many trials do we need so that information about the reward has arrived in the state just 'below' the starting state?

c. What happens to the learning speed if the number of states increases from 6 to 12? How many Q-values are non-zero after 3 trials? How many trial do we need so that information about the reward has arrived in the state just 'below' the starting state?

Solution:

a. In the first trial, since all Q's are zero, the term $(Q(s,a) - Q(s',a'))$ is always zero. Learning only occurs when there is a reward ie, the first time action a_1 is taken from state s_5. The learning is then

$$\Delta Q(s_5, a_1) = \eta \left[r - (Q(s_5,a_1) - Q(s_6,a_2)) \right] = \eta \,, \tag{14}$$

so that now all Q are zero except for $Q(s_5, a_1) = \eta$.

b. In the second trial, the first time $\Delta Q(s,a)$ is not zero is when the agent takes action a_1 from state s_4, and we have

$$\Delta Q(s_4, a_1) = \eta \left[r - (Q(s_4,a_1) - Q(s_5,a_1)) \right] = \eta(0 - (0 - \eta)) = \eta^2 \,. \tag{15}$$

Next, from state s_5, the agent chooses the action with the highest Q value, a_1, and the weight update is

$$\Delta Q(s_5, a_1) = \eta \left[r - (Q(s_5,a_1) - Q(s_6,a_2)) \right] = \eta(1 - (\eta - 0)) = \eta - \eta^2 \,. \tag{16}$$

So at the end of the second trial, the non-zero Qs are:

$$Q(s_4, a_1) = \eta^2 \qquad \text{and} \qquad Q(s_5, a_1) = 2\eta - \eta^2 \,.$$

In the third trial, the first Q update happens for $Q(s, a_1)$

$$\Delta Q(s, a_1) = \eta \left[r - (Q(s,a_1) - Q(s_4,a_1)) \right] = \eta(0 - (0 - \eta^2)) = \eta^3 \,. \tag{17}$$

The subsequent updates are

$$\begin{aligned}
\Delta Q(s_4, a_1) &= \eta \left[r - (Q(s_4,a_1) - Q(s_5,a_1)) \right] = \eta(0 - (\eta^2 - 2\eta + \eta^2)) = 2(\eta^2 - \eta^3) \\
\Delta Q(s_5, a_1) &= \eta \left[r - (Q(s_5,a_1) - Q(s_6,a_2)) \right] = \eta(1 - (2\eta - \eta^2 - 0)) = \eta - 2\eta^2 + \eta^3 \,.
\end{aligned}$$

So after three trials, the Qs are:

$$Q(s_3, a_1) = \eta^3 \,, \qquad Q(s_4, a_1) = 3\eta^2 - 2\eta^3 \qquad \text{and} \qquad Q(s_5, a_1) = 3\eta - 3\eta^2 + \eta^3 \,.$$

Note that terms for all the Qs converge towards 1 (the reward after). The higher η is, the faster the convergence, with convergence in 1 step in the extreme case $\eta = 1$.

We need one more trial until $Q(s_2, a_1)$ becomes non-zero, i.e. in total 4 trials.

c. Also with 12 states only 3 Q-values are non-zero after 3 trials. It takes 10 trials until the reward has arrived just 'below' the starting state.

Exercise 4. Bellman equation

Use the Bellman equation to calculate $Q(s, a1)$ and $Q(s, a2)$ for the scenario shown in Figure 1. Consider two different policies:

- Total exploration: All actions are chosen with equal probability.
- Greedy exploitation: The agent always chooses the best action.

Note that the rewards/next states are stochastic for the actions $a1'$, $a2'$ and $a3'$. Assume that the probabilities for the outcome of these actions are all equal.

Solution:

Total exploration: Start by computing the state-action values for states s'_1 and s'_2:

$$\begin{aligned} Q(s'_1, a'_1) &= \frac{1}{2}(1+2) = \frac{3}{2}\,, \\ Q(s'_1, a'_2) &= \frac{1}{2}(1+2) = \frac{3}{2}\,, \\ Q(s'_2, a'_2) &= \frac{1}{2}(1+2) = \frac{3}{2} \qquad \text{and} \\ Q(s'_2, a'_3) &= \frac{1}{2}(0+6) = 3. \end{aligned}$$

We can now compute the state-action values for state s:

$$\begin{aligned} Q(s, a_1) &= 1 + \frac{1}{2}(Q(s'_1, a'_1) + Q(s'_1, a'_2)) = \frac{5}{2} \qquad \text{and} \\ Q(s, a_2) &= 0 + \frac{1}{2}(Q(s'_2, a'_2) + Q(s'_2, a'_3)) = \frac{9}{4}\,. \end{aligned}$$

Greedy exploitation: In that case, the state-action values for the s'_1 and s'_2 are unchanged, but those for s reflect the fact that we now take the best action:

$$\begin{aligned} Q(s, a_1) &= 1 + Q(s'_1, a'_1) = \frac{5}{2} \qquad \text{and} \\ Q(s, a_2) &= 0 + Q(s'_2, a'_3) = 3\,. \end{aligned}$$

Notice that now the "best" action in state s is a_2, whereas it was a_1 for the total exploration policy.

Exercise 5. 3-step SARSA algorithm

In class we have discussed the SARSA algorithm (for discount factor $\gamma \leq 1$) and shown that, after convergence, the resulting Q-values solve (in expectation) the Bellman equation for *neighboring* states. Your friend claims that a 3-step SARSA for

$$\Delta Q(s_t, a_t) = \eta \left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 Q(s_{t+3}, a_{t+3}) - Q(s_t, a_t) \right]\,, \tag{18}$$

should work just as well.

To simplify the analysis, we assume that the environment has no loops (i.e., the graph is directed) so that we can consider $\gamma = 1$.

a. Assume that the 3-step SARSA algorithm converges in expectation. Proceed as during the lecture to show that $\langle \Delta Q(s_t, a_t) \rangle = 0$ implies

$$Q(s_t, a_t) = \sum_{s'} p^{a_t}_{s_t \to s'} \left[R^{a_t}_{s_t \to s'} + \sum_{a'} \pi(s', a') B_1(s', a') \right]$$

where

$$B_1(s',a') = \sum_{s''} p^{a'}_{s'\to s''}\left[R^{a'}_{s'\to s''} + \sum_{a''}\pi(s'',a'')B_2(s'',a'')\right]$$

$$B_2(s'',a'') = \sum_{s'''} p^{a''}_{s''\to s'''}\left[R^{a''}_{s''\to s'''} + \sum_{a'''}\pi(s''',a''')Q(s''',a''')\right]$$

b. Show the equivalence of the previous equation to the 1-step Bellman equation.

Solution:

a. $\langle \Delta Q(s_t,a_t)\rangle = 0$ implies

$$Q(s_t,a_t) = \sum_{s'} p^{a_t}_{s_t\to s'} R^{a_t}_{s_t\to s'} + \langle \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 Q(s_{t+3},a_{t+3})\rangle_{s_{t+1}=s'} \tag{19}$$

$$= \sum_{s'} p^{a_t}_{s_t\to s'}\left(R^{a_t}_{s_t\to s'} + \gamma\sum_{a'}\pi(s',a')\left(\sum_{s''} p^{a'}_{s'\to s''}R^{a'}_{s'\to s''} + \langle \gamma r_{t+2} + \gamma^2 Q(s_{t+3},a_{t+3})\rangle_{s_{t+2}=s''}\right)\right) \tag{20}$$

where we used the fact that the probability of action a' given state s' is determined by the policy $\pi(s',a')$. Continuing along the same lines we find the claimed result.

b. If we expand the 1-step Bellman equation by iteratively inserting the right-hand-side of the Bellman equation, we arrive at the 3 equations in a.