Solutions: SARSA and eligibility traces

Exercise 1 (in class): Iterative update

This is very similar to one of last week’s exercises: we define $\Delta Q_k$ as the difference between $Q_{k-1}$ and $Q_k$, and we simplify:

$$\Delta Q_k = Q_k - Q_{k-1} = \frac{1}{k} \sum_{i=1}^{k} r_i - \frac{1}{k-1} \sum_{i=1}^{k-1} r_i$$

$$= \frac{1}{k} \left( r_k + \sum_{i=1}^{k-1} r_i \right) - \frac{1}{k-1} \sum_{i=1}^{k-1} r_i$$

$$= \frac{1}{k} \left( r_k + \frac{k-1}{k-1} \sum_{i=1}^{k-1} r_i - \frac{k}{k-1} \sum_{i=1}^{k-1} r_i \right)$$

$$= \frac{1}{k} \left( r_k - \frac{1}{k-1} \sum_{i=1}^{k-1} r_i \right)$$

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

where we identified $\eta = 1/k$.

Exercise 2: Greedy policy and the two-armed bandit

2.1. 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 $r = 0.7$, its Q-value is updated to:

$$Q(a_2, t = 1) = Q(a_2, t = 0) + \Delta Q(a_2) = 0 + \eta (r - Q(a_2, t = 0)) = 0 + 0.2 \cdot 0.7 = 0.14.$$  

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

2.2. 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 $E[r_2] = 0.5 \cdot 0.7 + 0.5 \cdot 0 = 0.35$. The second action yields a higher reward on average.
2.3. Similarly as in 2.1., we can compute the Q-values after the first step with $\eta = 0.2$.
We obtain: $Q^*(a_1) = 1.8$ and $Q^*(a_2) = 1.74$. The online update rule can be transformed into a differential equation for the expected value of $Q$ if $\eta \ll 1$:

$$\Delta Q(a_i, t) = \eta (r_t - Q_{t-1})$$

$$\iff \frac{dE[Q(a_i, t)]}{dt} = E[r_i(t)] - E[Q(a_i, t)]$$

where we identified $\eta \approx dt$. We know that $E[r_1(t)] = E[r_1] = 0.25$ for $a_1$ and $E[r_2(t)] = E[r_2] = 0.35$ for $a_2$. The solution to the two differential equations is given by:

$$E[Q(a_1, t)] = (Q^*(a_1) - E[r_1]) \exp(-t) + E[r_1]$$

$$E[Q(a_2, t)] = (Q^*(a_2) - E[r_2]) \exp(-t) + E[r_2]$$

Initially, $E[Q(a_1, t)]$ will be higher than $E[Q(a_2, t)]$ although action $a_2$ has a higher expected reward. We therefore calculate the time $t$ at which the two curves cross such that the Q-value becomes higher for the best action:

$$E[Q(a_1, t)] = E[Q(a_2, t)] \rightarrow (Q^*(a_1) - E[r_1]) \exp(-t) + E[r_1] = (Q^*(a_2) - E[r_2]) \exp(-t) + E[r_2]$$

$$\rightarrow (1.8 - 0.25) \exp(-t) + 0.25 = (1.74 - 0.35) \exp(-t) + 0.35$$

$$\rightarrow (1.8 - 0.25 - 1.74 + 0.35) \exp(-t) = 0.35 - 0.25$$

$$\rightarrow (0.06 + 0.1) \exp(-t) = 0.1$$

$$\rightarrow t = -\log \frac{0.1}{0.16}$$

$$\rightarrow t \approx 0.47.$$ (5)

This corresponds to about 470 time steps (since $dt = \eta = 0.001$).

**Exercise 3: Bellman equation**

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

$$Q(s'_1, a'_1) = \frac{1}{2}(1 + 2) = \frac{3}{2},$$

$$Q(s'_1, a'_2) = \frac{1}{2}(1 + 2) = \frac{3}{2},$$

$$Q(s'_2, a'_2) = \frac{1}{2}(1 + 2) = \frac{3}{2} \quad \text{and}$$

$$Q(s'_2, a'_3) = \frac{1}{2}(0 + 6) = 3.$$

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

$$Q(s, a_1) = 1 + \frac{1}{2}(Q(s'_1, a'_1) + Q(s'_1, a'_2)) = \frac{5}{2} \quad \text{and}$$

$$Q(s, a_2) = 0 + \frac{1}{2}(Q(s'_2, a'_2) + Q(s'_2, a'_3)) = \frac{9}{4}.$$
**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:

$$Q(s, a_1) = 1 + Q(s'_1, a'_1) = \frac{5}{2} \quad \text{and} \quad Q(s, a_2) = 0 + Q(s'_2, a'_2) = 3.$$ 

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

**Exercise 4: SARSA algorithm**

4.1.: 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, i.e., the first time action $a_1$ is taken from state $s''$. The learning is then

$$\Delta Q(s'', a_1) = \eta \left[r - (Q(s'', a_1) - Q(s'', a_2))\right] = \eta,$$  

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

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

$$\Delta Q(s', a_1) = \eta \left[r - (Q(s', a_1) - Q(s'', a_1))\right] = \eta(0 - (\eta - 0)) = \eta^2.$$ 

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

$$\Delta Q(s'', a_1) = \eta \left[r - (Q(s'', a_1) - Q(s'', a_2))\right] = \eta(1 - (\eta - 0)) = \eta - \eta^2.$$ 

So at the end of the second trial, the non-zero $Q$s are:

$$Q(s', a_1) = \eta^2 \quad \text{and} \quad Q(s'', a_1) = 2\eta - \eta^2.$$ 

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

$$\Delta Q(s, a_1) = \eta \left[r - (Q(s, a_1) - Q(s', a_1))\right] = \eta(0 - (0 - \eta^2)) = \eta^3.$$ 

The subsequent updates are

$$\Delta Q(s', a_1) = \eta \left[r - (Q(s', a_1) - Q(s'', a_1))\right] = \eta(0 - (\eta^2 - 2\eta + \eta^2)) = 2(\eta^2 - \eta^3)$$

$$\Delta Q(s'', a_1) = \eta \left[r - (Q(s'', a_1) - Q(s'', a_2))\right] = \eta(1 - (2\eta - \eta^2 - 0)) = \eta - 2\eta^2 + \eta^3.$$ 

So after three trials, the $Q$s are:

$$Q(s, a_1) = \eta^3,$$  

$$Q(s', a_1) = 3\eta^2 - 2\eta^3 \quad \text{and} \quad Q(s'', a_1) = 3\eta - 3\eta^2 + \eta^3.$$ 

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