Tutorial 07 — Answers

(Version 1.1)

1. The action/value estimates at each point in time are as follows:

   1:
   - Action a_1 is selected, and payoff 9 is received, so:

$$Q_1(a_1) = 9$$

   - Action a_2 has not been selected, so $Q_1(a_2) = 0$

   2:
   - Action a_1 is selected, and payoff 2 is received, so, applying the formula on slide 17 of Lecture Reinforcement Learning 1:

$$\begin{aligned} Q_2(a_1) &= Q_1(a_1) + \frac{1}{2}(2 - Q_1(a_1)) \\ &= 9 + \left(\frac{-7}{2}\right) \\ &= 9 - 3.5 \\ &= 5.5 \end{aligned}$$

   - Action a_2 has not been selected, so $Q_2(a_2) = 0$

   3:
   - Action a_1 has not been selected, so $Q_3(a_1) = 5.5$
   - Action a_2 is selected and payoff 7 is received, so:

$$Q_3(a_2) = 7$$

   4:
   - Action a_1 has not been selected, so $Q_4(a_1) = 5.5$
   - Action a_2 is selected and payoff 3 is received, so:

$$Q_4(a_2) = 5$$

   doing exactly the same kind of calculation as for time 2.

   5:
   - Action a_1 is selected, and payoff 3 is received, so:

$$\begin{aligned} Q_5(a_1) &= Q_4(a_1) + \frac{1}{3}(3 - Q_4(a_1)) \\ &= 5.5 + \left(\frac{-2.5}{3}\right) \\ &= 4.67 \end{aligned}$$

   - Action a_2 has not been selected, so $Q_5(a_2) = 5$

   6:

$$\begin{aligned} Q_6(a_1) &= 4.67 \\ Q_6(a_2) &= 6 \end{aligned}$$

7:

$$Q_7(a_1) = 4.67$$
$$Q_7(a_2) = 5.5$$

8:

$$Q_8(a_1) = 4.75$$
$$Q_8(a_2) = 5.5$$

Note that, given the distribution from which the payoffs were drawn, we would expect the values to tend to 5, which they appear to be doing.

2. (a) ϵ-greedy picks the action with the greatest expected reward with probability $1 - \epsilon$, and picks randomly between the other actions with probability ϵ.

   Since a_2 has the greatest expected reward, a_2 will be picked with probability 0.9. With probability 0.1, a random selection will be made between the set of actions. Since there are three of these, each will be picked with probability 0.033.

   Thus, overall, a_1 and a_3 will each be picked with probability 0.033, and a_2 will be picked with probability 0.933.

   (b) We have the set of actions $A = \{a_1, a_2, a_3\}$ with:

$$Q(a_1) = 5$$
$$Q(a_2) = 7$$
$$Q(a_3) = 4$$

Softmax using the Gibbs distribution sets the probability of action a_i to be:

$$P(a_i) = \frac{e^{\frac{Q(a_i)}{\tau}}}{\sum_{a_j \in A} e^{\frac{Q(a_j)}{\tau}}}$$

So with $\tau = 0.1$:

$$\begin{aligned} e^{\frac{Q(a_1)}{\tau}} &= e^{\frac{5}{0.1}} \\ &= 5.18 \times 10^{21} \\ e^{\frac{Q(a_2)}{\tau}} &= e^{\frac{7}{0.1}} \\ &= 2.5 \times 10^{30} \\ e^{\frac{Q(a_3)}{\tau}} &= e^{\frac{4}{0.1}} \\ &= 2.35 \times 10^{17} \end{aligned}$$

and:

$$P(a_1) \approx 0$$
$$P(a_2) \approx 1$$
$$P(a_3) \approx 0$$

and action selection is approximately greedy. However, if $\tau = 10$, we would have:

$$\begin{aligned} e^{\frac{Q(a_1)}{\tau}} &= e^{\frac{5}{10}} \\ &= 1.65 \\ e^{\frac{Q(a_2)}{\tau}} &= e^{\frac{7}{10}} \\ &= 2.01 \\ e^{\frac{Q(a_3)}{\tau}} &= e^{\frac{4}{10}} \\ &= 1.49 \end{aligned}$$

and

$$\begin{aligned} P(a_1) &= 0.32 \\ P(a_2) &= 0.39 \\ P(a_3) &= 0.29 \end{aligned}$$

3. Once these utility values are computed, the agent can establish a policy by picking the action with the greatest expected utility in each state.

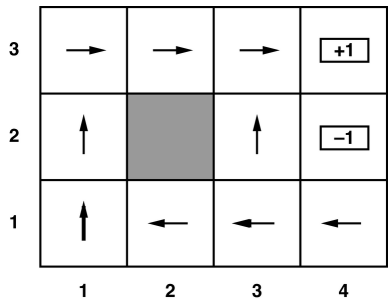
   For example, if the agent were to pick Up in state (1, 1), its expected utility would be:

$$\begin{aligned} EU(Up_{(1,1)}) &= 0.8 \times U((1,2)) + 0.1 \times U((1,1)) + 0.1 \times U((2,1)) \\ &= 0.8 \times 0.762 + 0.1 \times 0.705 + 0.1 \times 0.655 \\ &= 0.6096 + 0.0705 + 0.0655 \\ &= 0.7456 \end{aligned}$$

   Similarly:

$$\begin{aligned} EU(Left_{(1,1)}) &= 0.8 \times U((1,1)) + 0.1 \times U((1,1)) + 0.1 \times U((1,2)) \\ &= 0.711 \\ EU(Right_{(1,1)}) &= 0.8 \times U((2,1)) + 0.1 \times U((1,1)) + 0.1 \times U((1,2)) \\ &= 0.671 \\ EU(Down_{(1,1)}) &= 0.8 \times U((1,1)) + 0.1 \times U((1,1)) + 0.1 \times U((2,1)) \\ &= 0.7 \end{aligned}$$

   So the action with greatest expected utility in $(1, 1)$ is Up. The full set of actions are those in this (familiar) figure:



since the optimal policy is the result of picking the best (greedy) action in each state when you have the final utility values form value iteration (which are the utility values under the optimal policy).

Make sure you go through the calculation for $(3, 1)$ and understand why $Left$ is the action selected.

4. (a) After the first run, we have the following utilities.

   (1, 1) We have one sample which is 12 steps from the goal. The estimated utility of this state is thus

   $$U(1,1) = 1 - 12 \times 0.04$$
   $$= 0.52$$

   (1, 2) We have one sample, 11 steps from the goal. The estimated utility is thus 0.56

   (1, 3) We have two samples, 9 and 10 steps from the goal. The two samples are thus 0.6 and 0.64. The estimated utiluity of the state is the everage of these, 0.62.

   (2, 3) We have three samples, 6, 7 and 8 steps from the goal. The estimated utility is thus 0.72.

   (3, 3) Again we have three samples, 1, 3 and 5 steps from the goal, with correspondoing utility estimates of 0.96, 0.88 and 0.8. The average is 0.88.

   (3, 2) Two estimates, 2 and 4 steps from the goal, giving an estimated utility of 0.88.

   We have no data on the other states, so their estimated utility is zero.

   (b) $P((1,2)|(1,1), Up)$: The only time the agent tries to move Up in (1, 1), so the probability is 1.

   $P((1,3)|(1,2), Up)$: as in the previous case, this is 1.

   $P((2,3)|(1,3), Right)$: when the agent moves Right in (1, 3), it only gets to (2, 3) once in two attempts, so the probability is 0.5.

   $P((1,3)|(1,3), Right)$: the other time the agent tries to move Right in (1, 3) it ends up in the same state, so this is 0.5.

   $P((2,3)|(2,3), Right)$: two out of three times that the agent moves Right in (2, 3), it remains in the same state, so this probability is 0.667.

   $P((3,3)|(2,3), Right)$: 0.333 (what happens the other time that the agent tries to move Right in (2, 3)).

   $P((3,2)|(3,3), Right)$: 0.667

   $P((4,3)|(3,3), Right)$: 0.333

   $P((3,3)|(3,2), Up)$: The agent makes this move twice, and both times moves to (3, 3), so the probability is 1.

   All other transition probabilities are zero, because we have no samples.

   (c) The second run gives us more data on some of the states.

   (1, 1) We have a new sample, 6 steps from the goal, giving a value of 0.76. The new estimated values is the average of this and the previous sample, 0.64.

   (1, 2) We have two new samples, 4 and 5 steps from the goal, 0.84 and 0.8. Combining these with the previous estimate of 0.56, we get an (average) estimated reward of 0.73.

   (1, 3) One more sample, 0.88, to add to the previous estimate gives us 0.707.

   (2, 3) 0.77.

(3, 3) 0.9.

All other states have the same estimated utility as after the first run.

Updating the probabilities using the second run we have new values for:

$P((1,2)|(1,1),Up)$: still 100% success, so the probability is still 1.

$P((1,3)|(1,2),Up)$: now we have three attempts to move Up in (1, 2), two of which succeed, so the probability is 0.667.

$P((1,2)|(1,2),Up)$: because of the one case in which this does not succeed, this is 0.33.

$P((2,3)|(1,3),Right)$: a success in the second run means that this value is also 0.667.

$P((1,3)|(1,3),Right)$: 0.333.

$P((2,3)|(2,3),Right)$: moving right in (2, 3) now succeeds in moving the agent to (3, 3) 2 times out of 4, so this value is 0.5.

$P((3,3)|(2,3),Right)$: by the same reasoning, this is also 0.5.

$P((4,3)|(3,3),Right)$: 0.5

$P((3,2)|(3,3),Right)$: 0.5

All other values stay the same.

(d) Taking the third run into account we get:

(1, 1) 0.71

(1, 2) 0.75

(1, 3) 0.75

(2, 3) 0.8

(3, 3) 0.912

All other states have the same estimated utility as after the second run.

$P((1,2)|(1,1),Up)$: 0.75

$P((1,1)|(1,1),Up)$: 0.25

$P((1,3)|(1,2),Up)$: 0.75

$P((1,2)|(1,2),Up)$: 0.25

$P((2,3)|(1,3),Right)$: 0.75.

$P((1,3)|(1,3),Right)$: 0.25.

$P((2,3)|(2,3),Right)$: 0.4

$P((3,3)|(2,3),Right)$: 0.6

$P((4,3)|(3,3),Right)$: 0.6

$P((3,2)|(3,3),Right)$: 0.4

All other values stay the same.

(e) What we see is that the estimates all tend towards the correct values (which we know from Lecture 4), pretty quickly in some cases, though some values (like that for Right in (3, 3) are distorted by some unlikely sequences of events).

Version list

- Version 1.0, March 5th 2020.
- Version 1.1, February 5th 2021.