Reinforcement Learning
Prof. Milica Gašić

Release: Thursday, 7 November 2024
Discussion: Wednesday, 13 November 2024

Exercise set #5

You do not have to hand in your solutions to the exercises and they will **not** be graded. However, there will be four short tests during the semester. You need to reach ≥50% of the total points in order to be admitted to the final exam (Klausur). The tests are held at the start of a lecture (room 2522.U1.74) at the following dates:

Test 1: Thursday, 31 October 2024, 10:30-10:45
Test 2: Thursday, 21 November 2024, 10:30-10:45
Test 3: Thursday, 5 December 2024, 10:30-10:45
Test 4: Thursday, 9 January 2025, 10:30-10:45

Please ask questions in the RocketChat
The exercises are discussed every Wednesday, 14:30-16:00 in room 2512.02.33.

1. **TD and MC prediction**

   In the lecture we have seen *batch* versions of TD and MC prediction (lecture 6, slide 17). Assume an MDP with a single action a_0 that has no effect (also known as Markov reward process). Given the following batch of episodes (i.e., sequences $s_0, r_1, s_1, r_2, \ldots$, the action is a_0 in every step)

   (1) $A, 0, B, 6, C$
   (2) $A, 3, C$
   (3) $A, 2, B, 0, A, 3, C$
   (4) $B, 0, A, 2, B, 2, C$
   (5) $B, 0, A, 3, C$
   (6) $B, 2, C$
   (7) $B, 6, C$
   (8) $B, 2, C$
   (9) $B, 6, C$

   and assuming no discounting (i.e., $\gamma = 1$), your tasks are the following:

   (a) Apply batch TD to the episodes (1), (2), (3), i.e., apply Algorithm 2 from lecture 6, where the tuples are sampled from the batch in chronological order. Initialize all values with zero and use the learning rate $\alpha = 0.1$.

   Answer: Simplify the update rule:

   $$V(s_t) = V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t))$$
   $$V(s_t) = (1 - \alpha)V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}))$$

   For $\gamma = 1, \alpha = 0.1$:

   $$V(s_t) = 0.9 \cdot V(s_t) + 0.1(r_{t+1} + V(s_{t+1}))$$

Exercises by Stefan Harmeling, used with permission

s_t	r_{t+1}	s_{t+1}	$V(A)$	$V(B)$
			0	0
A	0	B	$0.9 \cdot 0 + 0.1(0 + 0) = 0$	0
B	6	C	0	$0.9 \cdot 0 + 0.1(6 + 0) = 0.6$
A	3	C	$0.9 \cdot 0 + 0.1(3 + 0) = 0.3$	0.6
A	2	B	$0.9 \cdot 0.3 + 0.1(2 + 0.6) = 0.53$	0.6
B	0	A	0.53	$0.9 \cdot 0.6 + 0.1(0 + 0.56) = 0.596$
A	3	C	$0.9 \cdot 0.53 + 0.1(3 + 0) = 0.777$	0.596

(b) Derive the MDP that best fits the data, i.e., calculate the most likely transition probabilities $\mathcal{P}(s'|s,a)$ and reward function $\mathcal{R}(s,a)$, and draw the graphical model.

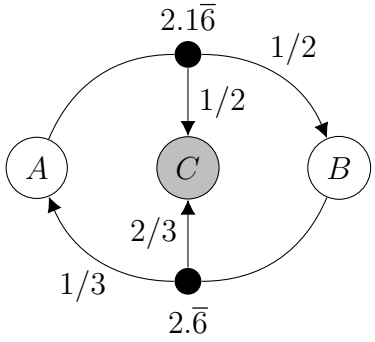
Answer: From lecture 6, slide 19 we know:

$$\mathcal{P}(s'|s,a) = \frac{1}{N(s,a)} \sum_{k=1}^{K} \sum_{t=0}^{T_k-1} [s_{t+1}^k = s'][s_t^k = s][a_t^k = a] = \frac{N(s,a,s')}{N(s,a)}$$

$$\mathcal{R}(s,a) = \frac{1}{N(s,a)} \sum_{k=1}^{K} \sum_{t=0}^{T_k-1} [s_t^k = s][a_t^k = a] r_{t+1}^k = \frac{\text{sum of rewards in } (s,a)}{N(s,a)}$$

State-action pairs:
$(A, a_0) : 6$
$(B, a_0) : 9$
Transitions:
$(A, a_0, A) : 0$
$(A, a_0, B) : 3$
$(A, a_0, C) : 3$
$(B, a_0, A) : 3$
$(B, a_0, B) : 0$
$(B, a_0, C) : 6$
Rewards:
$(A, a_0) : 0 + 3 + 2 + 3 + 2 + 3 = 13$
$(B, a_0) : 6 + 0 + 0 + 2 + 0 + 2 + 6 + 2 + 6 = 24$

$$\begin{aligned}
\Rightarrow \mathcal{P}(A|A,a_0) &= 0/6 = 0 \\
\mathcal{P}(B|A,a_0) &= 3/6 = 1/2 \\
\mathcal{P}(C|A,a_0) &= 3/6 = 1/2 \\
\mathcal{P}(A|B,a_0) &= 3/9 = 1/3 \\
\mathcal{P}(B|B,a_0) &= 0/9 = 0 \\
\mathcal{P}(C|B,a_0) &= 6/9 = 2/3 \\
\mathcal{R}(A,a_0) &= 13/6 = 2.1\overline{6} \\
\mathcal{R}(B,a_0) &= 24/9 = 2.\overline{6}
\end{aligned}$$



(c) Calculate the values for states A and B using the MDP derived in part (b). Since there is only one action, the policy does not matter. From the lecture we know that batch TD converges to this solution.

Answer:

$$
\begin{aligned}
V(A) &= \tfrac{13}{6} + \tfrac{1}{2}V(B) + \tfrac{1}{2}V(C) = \tfrac{13}{6} + \tfrac{1}{2}V(B) \\
V(B) &= \tfrac{24}{9} + \tfrac{1}{3}V(A) + \tfrac{2}{3}V(C) = \tfrac{24}{9} + \tfrac{1}{3}V(A) \\
\Rightarrow V(B) &= \tfrac{24}{9} + \tfrac{1}{3}\left(\tfrac{13}{6} + \tfrac{1}{2}V(B)\right) \\
&= \tfrac{61}{18} + \tfrac{1}{6}V(B) \\
\Leftrightarrow \left(1 - \tfrac{1}{6}\right)V(B) &= \tfrac{61}{18} \\
\Leftrightarrow V(B) &= \tfrac{61}{15} = 4.0\overline{6} \\
\Rightarrow V(A) &= \tfrac{13}{6} + \tfrac{1}{2} \cdot \tfrac{61}{15} = \tfrac{21}{5} = 4.2
\end{aligned}
$$

(d) Calculate the values for states A and B by applying batch first-visit and every-visit MC to the given episodes.

Answer:
Calculate returns:

(1) $A, 6, B, 6, C$
(2) $A, 3, C$
(3) $A, 5, B, 3, A, 3, C$
(4) $B, 4, A, 4, B, 2, C$
(5) $B, 3, A, 3, C$
(6) $B, 2, C$
(7) $B, 6, C$
(8) $B, 2, C$
(9) $B, 6, C$

First-visit:

$$
\begin{aligned}
V(A) &= \frac{6+3+5+4+3}{5} = 4.2 \\
V(B) &= \frac{6+3+4+3+2+6+2+6}{8} = 4
\end{aligned}
$$

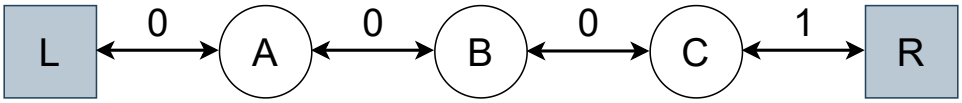
Every-visit:

$$
\begin{aligned}
V(A) &= \frac{6+3+5+3+4+3}{6} = 4 \\
V(B) &= \frac{6+3+4+2+3+2+6+2+6}{9} = \tfrac{34}{9} = 3.\overline{7}
\end{aligned}
$$

2. **TD and MC prediction**

A Markov reward process can be viewed as a Markov decision process with exactly one action a_0 that has no influence on the reward and dynamics. The following graph depicts

an example of a Markov reward process (MRP). In this MRP, every episode starts in the center state B, then proceed either left or right by one state on each step, with equal probability. Episodes terminate either on the extreme left or right in states L or R. When an episode terminates in R, a reward of 1 occurs; all other rewards are 0.



(a) Your friend guesses that the state-values for the MRP are given as $v(L) = 0, v(A) = 1/4, v(B) = 1/2, v(C) = 3/4, v(R) = 0$. Proof that these are the true values.

Answer: The true state-value function v_π is the unique solution to the Bellman expectation equation

$$v_\pi(s) = \sum_a \pi(a|s) \cdot [R(s,a) + \sum_{s'} P(s'|s,a) v_\pi(s')].$$

We only have a single action a_0 that has no effect on the process, so we omit it in the following. $v(L) = v(R) = 0$ holds because these are terminal states. We have $R(A) = R(B) = 0.5 \cdot 0 + 0.5 \cdot 0 = 0$ and $R(C) = 0.5 \cdot 0 + 0.5 \cdot 1 = 0.5$. It follows that

$$R(A) + 0.5 \cdot v(L) + 0.5 \cdot v(B) = 0 + 0.5 \cdot 0 + 0.5 \cdot 1/2 = 1/4 = v(A)$$
$$R(B) + 0.5 \cdot v(A) + 0.5 \cdot v(C) = 0 + 0.5 \cdot 1/4 + 0.5 \cdot 3/4 = 1/8 + 3/8 = 1/2 = v(B)$$
$$R(C) + 0.5 \cdot v(B) + 0.5 \cdot v(R) = 0.5 + 0.5 \cdot 1/2 + 0.5 \cdot 0 = 1/2 + 1/4 = 3/4 = v(C)$$

(b) Name two algorithms that would also lead to the true state-value function.

Answer: Policy evaluation using Bellman expectation equation, Monte-Carlo prediction, TD-prediction

(c) Use Monte-Carlo prediction and TD-prediction in order to produce estimates for $v(A), v(B), v(C)$ by generating 3 episodes using the Markov reward process.