

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 $\geq 40\%$ 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 25.22.U1.74) at the following dates:

Test 1: Thursday, 6 November 2025, 10:30-10:45

Test 2: Thursday, 27 November 2025, 10:30-10:45

Test 3: Thursday, 11 December 2025, 10:30-10:45

Test 4: Thursday, 15 January 2026, 10:30-10:45

Please ask questions in the RocketChat

The exercises are discussed every Wednesday, 14:30-16:00 in room 25.12.02.33.

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 or MRP). Given the following batch of episodes (i.e., sequences $s_0, r_1, s_1, r_2, \dots$, the action is a_0 in every step)

(1) $A, 0, B, 6, C$

(6) $B, 2, C$

(2) $A, 3, C$

(7) $B, 6, C$

(3) $A, 2, B, 0, A, 3, C$

(8) $B, 2, C$

(4) $B, 0, A, 2, B, 2, C$

(9) $B, 6, C$

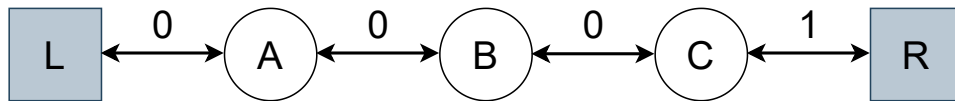
(5) $B, 0, A, 3, C$

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

- 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 $\alpha = 0.1$.
- 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.
- 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.
- Calculate the values for states A and B by applying batch first-visit and every-visit MC to the given episodes.

2. TD and MC prediction

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.
- (b) Name two algorithms that would also lead to the true state-value function.
- (c) Implement Monte-Carlo prediction and TD-prediction in order to produce estimates for $v(A), v(B), v(C)$ of the Markov reward process. Follow the instructions in the Jupyter notebook `mc-td-prediction.ipynb`.