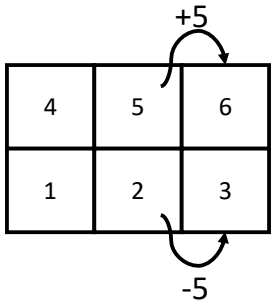
Reinforcement Learning (SS18) - Exercise 8

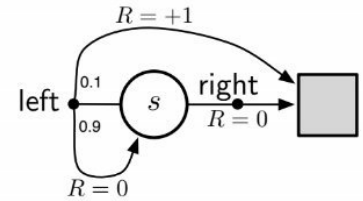
Daniel Hennes

no points / no due date

This gridworld MDP operates like to the one we saw in class. There are two terminal goal states (3 and 6) with reward +5 and and −5 for transitions to the terminal state. Rewards are 0 for all non-terminal transitions. The dynamics are such that the intended action (North, South, West, or East) happens with probability 0.8. With probability 0.1 each, the agent ends up in one of the states perpendicular to the intended direction. If a collision with a wall happens, the agent stays in the previous state. The agent starts in state 1.



1. What is the optimal policy?
2. Suppose the agent knows the transition probabilities. Give the first two iterations of value iteration updates for each state, with a discount of $\gamma = 0.9$. (Assume V_0 is 0 everywhere and compute V_1 and V_2.)
3. Find the optimal value of state 1.
4. What are the Monte Carlo estimates for states 1 and 5 given these traces?
   - 1 – 2 – 3
   - 1 – 2 – 5 – 6
   - 1 – 4 – 5 – 6
5. Using a learning rate of $\alpha = 0.1$ and assuming initial values of 0, what updates does the TD-learning agent make after the first two trials given above?
6. Given the following MDP, a target policy π, and sample episodes generated from a behaviour policy b: $e_1 = (s, \rightarrow, 0), e_2 = (s, \leftarrow, 0, s, \leftarrow, 1), e_3 = (s \leftarrow, 0, s, \leftarrow, 0, s, \rightarrow, 0), e_4 = (s, \leftarrow, 1), e_5 = (s, \leftarrow, 0, s, \rightarrow, 0)$. Assuming discount factor $\gamma = 1$, calculate the Monte-Carlo value estimates for V_π using ordinary importance sampling. What are the true values?



$$\pi(\text{left}|s) = 1$$

$$b(\text{left}|s) = \frac{1}{2}$$

R. Sutton and A. G. Barto, Reinforcement Learning: An Introduction

7. Draw the backup diagram of 4-step Expected Sarsa and label all entities.

8. Write an algorithm for Linear Sarsa, that is Sarsa with linear approximation of the action-value function $Q(s, a; \boldsymbol{w}) = \boldsymbol{x}(s, a)^T \boldsymbol{w}$. Now assume Quadratic Sarsa with $Q(s, a; \boldsymbol{W}) = \boldsymbol{x}(s, a)^T \boldsymbol{W} \boldsymbol{x}(s, a)$. How does the algorithm change.

9. Show that:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \mathbb{E}_\tau[G_0] &= \mathbb{E}_\tau\Big[\sum_{t=0}^{T-1} \nabla_{\boldsymbol{\theta}} \log \pi(A_t \mid S_t, \boldsymbol{\theta}) \sum_{k=t+1}^{T} R_k\Big] \\ &= \mathbb{E}_{\pi_{\boldsymbol{\theta}}}\big[q_\pi(S_t, A_t) \nabla \log \pi(A_t \mid S_t, \boldsymbol{\theta})\big]\end{aligned}$$