Reinforcement Learning
Assignement - MDPs, Value Functions and DP

In this assignment we will use DP to compute value functions and optimize the policy of simple MDPs. This will allow us to understand the basic concepts that support not only DP but also RL algorithms.

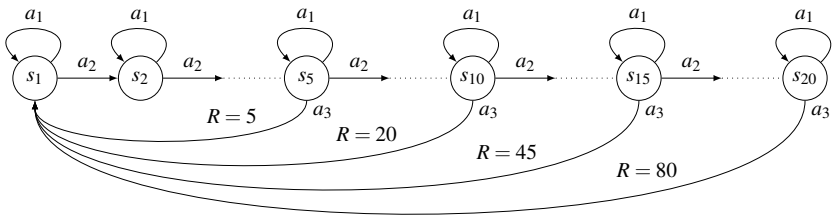

Figure 1: Cycles MDP. There are twenty states in a row and three actions: a_1 leaves the state unchanged; a_2 leads to the state to the right, except for s_{20}, in which a_2 does not exist; a_3 only exists in states $s_5, s_{10}, s_{15}, s_{20}$ and leads to s_1. The reward is always 0 except when taking action a_3.

1. **A simple MDP with deterministic dynamics**

   Consider the simple "Cycles" MDP depicted in Figure 1.

   Consider two deterministic policies: π_1 always selects a_2, except when in $s_5, s_{10}, s_{15}, s_{20}$, where it selects a_3; π_2 always selects a_2, except when in s_{20}, where it selects a_3. In the following, we will consider different discount factors $\gamma \in \{0.5, 0.85, 0.9, 0.99, 1 - 10^{-5}, 1\}$.

   (a) Compute the exact value of V^{π_1}, V^{π_2}. Verify that you obtained the correct solution by evaluating the Bellman expectation equation. What do you observe when the discount factor changes?

   Hint: since the MDP has few discrete states and actions you can use the explicit formula with cubic complexity.

   (b) Compute the exact action-value functions Q^{π_1}, Q^{π_2}. Check that $Q^{\pi_i}(s, \pi_i(s)) = V^{\pi_i}(s)$. Are these policies optimal? What do you observe when the discount factor changes?

   (c) Compute V^{π_i} by Policy Evaluation. How many iterates do you need to obtain a good accuracy? In order to investigate the convergence of the algorithm, plot the norm of the error with respect to the exact solution and of the update at each iteration for each value of γ.

   (d) Compute $V^\star$ by Value Iteration

   (e) Compute $V^\star$ by Policy Evaluation + Policy Improvement

   (f) **Optional** What are τ_{π_i} and $\tau^\gamma_{\pi_i}$ if the initial state is s_{16}? And if it is s_1?

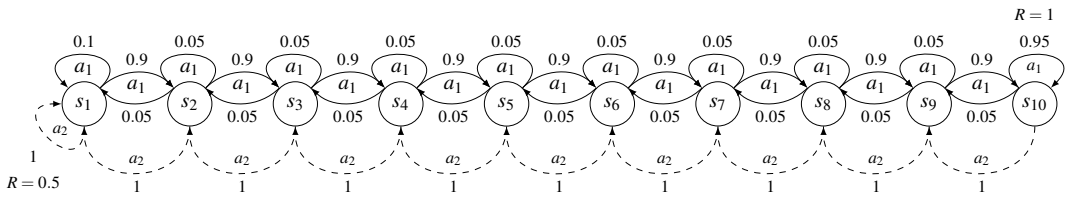

Figure 2: RiverSwim MDP. There are ten states in a row and two actions: a_1 leads to the state to the right with probability 0.9, to the same state with probability 0.05 and to the state to the left with probability 0.05; a_2 always leads to the state to the left. The reward is always 0, except $R(s_{10},a_1)=1$ and $R(s_1,a_2)=0.5$.

2. **A simple MDP with stochastic dynamics**

   Consider the simple "River Swim" MDP depicted in Figure 2.

   (a) Consider a stochastic policy π which selects a_1 with probability p and action a_2 with probability $1-p$. In the following, we will consider discount factor $\gamma = 0.99$. Compute the exact value of V^π. Verify that you obtained the correct solution by evaluating the Bellman expectation equation. What do you observe when p changes?

   (b) Compute $V^\star$ by Value Iteration for $\gamma = 0.99$. What is the optimal policy? Why? What if we select $\gamma = 0.9$?

   (c) **Optional** For $\gamma = 0.99$, what are $\tau_{\pi_\star}$ and $\tau^\gamma_{\pi_\star}$ if the initial state is s_1?

3. **A simple LQR: Optional**

   Consider a very simple LQR problem, defined as

$$s_+ = \boldsymbol{A}s + \boldsymbol{B}a + w \qquad\qquad R(s,a) = s^\top \boldsymbol{Q}s + a^\top \boldsymbol{R}a,$$

   with $w \sim \mathcal{N}\left(0, 10^{-4}\right)$, $\boldsymbol{A} = 1.1$, $\boldsymbol{B} = 1$, $\boldsymbol{Q} = 0.5$, $\boldsymbol{R} = 1$, and discount factor $\gamma = 0.9$.

   (a) **Optional** What are the optimal policy and value function?
   Hint: A discounted LQR problem can be solved as the undiscounted LQR problem with system matrices $\sqrt{\gamma}\boldsymbol{A}$, $\sqrt{\gamma}\boldsymbol{B}$. You should be able to prove that using the DARE.

   (b) **Optional** Solve the Policy Evaluation problem for $\boldsymbol{K} = 1$.

   (c) **Optional** Solve the problem by the Value Itaration.

   (d) **Optional** Solve the problem by the Policy Itaration.

   (e) **Optional** What are $\tau_{\pi_\star}$ and $\tau^\gamma_{\pi_\star}$ if the initial state is $s = 0$?