Prof. Dr. Leif Döring
André Ferdinand, Sara Klein

Reinforcement Learning

6. Excercise Sheet

1. Policy evaluation

Consider Algorithm 8 from the lecture. In Theorem 3.3.2 we proved convergence for this algorithm if $\gamma < 1$. Now assume $\gamma = 1$ and set $\Delta = 2\epsilon$ in the initialisation and choose termination condition $\Delta < \epsilon$. Give an example such that Algorithm 8 does not converge using $\gamma = 1$. Your are allowed to initialise the value function V arbitrarily.

2. Convergence of the in-place policy evaluation algorithm

Recall Algorithm 9 from the lecture. We aim to prove convergence of the algorithm (without termination) to V^π. Therfore label the state space $\mathcal{S}$ by $s_1, \dots, s_K$ and define

$$T_s^\pi V(s') = \begin{cases} T^\pi V(s) & : s = s' \\ V(s) & : s \neq s' \end{cases}$$

Define the composition $\overline{T}^\pi : U \to U$, $\overline{T}^\pi(v) := \left(T^\pi_{s_K} \circ \dots \circ T^\pi_{s_1}\right)(v)$ on the space of all functions $U = \{u : \mathcal{S} \to \mathbb{R}\}$ equiped with the supremums norm.

a) Argue why $\overline{T}^\pi$ is different from the Bellman operator T^π.

b) Show that V^π is a fixpoint of the operator $\overline{T}^\pi$.

c) Prove that $\overline{T}^\pi$ is a contraction on $(U, ||\cdot||_\infty)$.

3. Monte Carlo generalised ε-greedy policy iteration

In this exercise, we will use a Monte Carlo estimator to perform policy iteration. After Grid World has been a first example for a Markov decision process, we will now deal with more complex examples.

a) Implement the ice vendor (Example 3.1.8) as a Markov decision process.

b) Use the policy iteration from the last lecture to get the optimal decision rule for the iceman.

c) Implement the Monte Carlo generalised ε-greedy policy iteration (algorithm 18) to get the optimal decision rule for the iceman. Additionally, use different ε parameters as well as a sequence $\varepsilon_n \downarrow 0, n \to \infty$.