Machine Learning, Tutorial 12
Universität Bern

Simon Jenni (jenni@inf.unibe.ch)

Reinforcement Learning

1. Prove that adding a constant C to all the rewards in a deterministic MDP will add a constant K to the values of all the states. This therefore does not affect the relative values of states under any policies. What is K in terms of C and γ?

   Solution. Since the MDP is deterministic the Bellman equation has the form

   $$V^{\pi}(s) = R(s) + \gamma V^{\pi}(f_{\pi}(s))$$

   where $f_{\pi} : S \to S$ is the transition function when the policy π is applied.

   The above formula can be expanded,

   $$V^{\pi}(s) = R(s) + \gamma R(f_{\pi}(s)) + \gamma^2 V^{\pi}(f_{\pi}(f_{\pi}(s))) = \sum_{k=0}^{\infty} \gamma^k R(f_{\pi}^{(k)}(s))$$

   where $f^{(k)}(s)$ means that we apply the function f on the variables, k times. If we add a constant C to all rewards, the value becomes,

   $$V_C^{\pi}(s) = \sum_{k=0}^{\infty} \gamma^k (R(f_{\pi}^{(k)}(s)) + C) = \sum_{k=0}^{\infty} \gamma^k R(f_{\pi}^{(k)}(s)) + C \sum_{k=0}^{\infty} \gamma^k = V^{\pi}(s) + \frac{C}{1-\gamma}$$

   So $K = \frac{C}{1-\gamma}$. In the derivation above we used the fact that $\sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$ when $0 < \gamma < 1$.

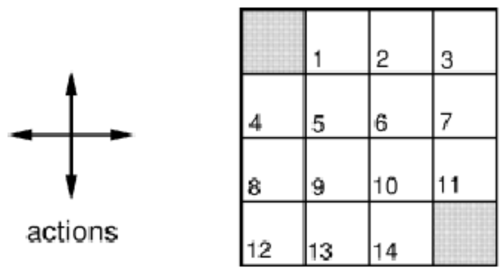

Figure 1: Gridworld illustration with available actions (*left*) and numbered nonterminal state locations (*right*). The terminal state locations are greyed out.

2. Consider the gridworld shown in Fig. 1 with a Markov decision process $M = (\mathcal{S}_N \cup \mathcal{S}_T, \mathcal{A}, \mathcal{P}_{sa}, \gamma, \mathcal{R})$. The nonterminal states are $\mathcal{S}_N = \{1, 2, \ldots, 14\}$ and the terminal states $\mathcal{S}_T$ are shaded in the figure. There are four possible actions in each non-terminal state, $\mathcal{A} = \{\text{up}, \text{down}, \text{right}, \text{left}\}$. Transitions that would take the agent off the grid in fact leave the state unchanged. The reward is 0 for a terminal state and $\mathcal{R}(s) = -1$ for all non-terminal states. The discount factor is set to $\gamma = 0.5$ and the state transition probabilities are given by:

$$\mathcal{P}_{sa}(s') = \begin{cases} 0.7 & \text{if } a(s) = s' \\ 0.1 & \text{if } a(s) \neq s' \text{ and } \exists a' \in \mathcal{A} \text{ with } a'(s) = s' \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Suppose that we use **value iteration** with synchronous update to compute the value function. Fig. 2 shows the sequence of value functions computed by the value iteration algorithm (k is the iteration index).

Fill the blank cells in the state value table for iterations $k = 2$ and $k = 3$ in Fig. 2. Show all your calculations.

Hint: The value function update is given by

$$V(s) := \mathcal{R}(s) + \gamma \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}_N \cup \mathcal{S}_T} \mathcal{P}_{sa}(s') V(s'). \tag{2}$$

k=0

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0

k=1

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0

k=2

0.0		-1.5	-1.5
	-1.5		-1.5
-1.5	-1.5	-1.5	
-1.5	-1.5		0.0

k=3

0.0		-1.63	-1.75
	-1.61		-1.63
-1.63	-1.75	-1.61	
-1.75	-1.63		0.0

Figure 2: The first three iterations of the value iteration algorithm on the gridworld example. Fill in the missing values.

k=0

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0

k=1

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0

k=2

0.0	-1.15	-1.5	-1.5
-1.15	-1.5	-1.5	-1.5
-1.5	-1.5	-1.5	-1.15
-1.5	-1.5	-1.15	0.0

k=3

0.0	-1.21	-1.63	-1.75
-1.21	-1.61	-1.75	-1.63
-1.63	-1.75	-1.61	-1.21
-1.75	-1.63	-1.21	0.0

Figure 3: **Solution.** Example calculation for state 1 ($k = 2$): $V(1) = \mathcal{R}(1) + \gamma \sum_{s' \in \mathcal{S}_N \cup \mathcal{S}_T} \mathcal{P}_{1,left}(s') V(s') = -1 + 0.5(0.7 \cdot 0.0 + 3 \cdot 0.1 \cdot (-1.0)) = -1.15.$