Exercise Sheet 1 - Solutions*

Omar D. Domingues
omar (dot) darwiche-domingues (at) inria.fr

Last update: January 18, 2019

1 Exercise 1

The value function in this case can be defined as:

$$V_\pi(x) = E_\pi \left[\sum_{t=0}^{\infty} r(X_t, A_t, X_{t+1}) \Big| X_0 = x \right] \tag{1}$$

where $r(X_t, A_t, X_{t+1})$ is the reward obtained when the agent is at state X_t, takes action A_t and goes to state X_{t+1}. This sum is guaranteed to converge, since the MDP in this exercise has an absorbing state. We can rewrite $V_\pi(x)$ as:

$$\begin{aligned}
V_\pi(x) &= E_\pi \left[r(X_0, A_0, X_{t+1}) + \sum_{t=1}^{\infty} r(X_t, A_t, X_{t+1}) \Big| X_0 = x \right] \\
&= \sum_{x' \in \mathcal{S}, a \in \mathcal{A}} \pi(a|x) P(x'|s,a) r(x,a,x') + E_\pi \left[E_\pi \left[\sum_{t=1}^{\infty} r(X_t, A_t, X_{t+1}) \Big| X_1, X_0 = x \right] \Big| X_0 = x \right] \\
&= \sum_{x' \in \mathcal{S}, a \in \mathcal{A}} \pi(a|x) P(x'|x,a) r(x,a,x') + E_\pi \left[V_\pi(X_1) \Big| X_0 = x \right] \\
&= \sum_{x' \in \mathcal{S}, a \in \mathcal{A}} \pi(a|x) P(x'|x,a) \left[r(x,a,x') + V_\pi(x') \right]
\end{aligned} \tag{2}$$

In this exercice, we have $x \in \mathcal{S} = \{1, 2, 3\}$, $a \in \mathcal{A} = \{"a", "b"\}$, and $r(x, a, x') = \mathbb{I}\{x = 2, a = "a", x' = 2\}$. The policy $\pi(a|x)$ and the transition probabilities $P(x'|x, a)$ are given for all $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. We can then find $V_\pi(1), V_\pi(2), V_\pi(3)$ by solving the linear system of equations defined above.

2 Exercise 2

We have:

- The state space is given by $\mathcal{S} = \{1, \dots, N\}$, where N is the number of cities;
- The action space $\mathcal{A}$ is the set of actions available in each city (concretely it could be the set of directions or roads to take);
- The transition probabilities $\tilde{p}(j|i, a)$, which represent the probability of getting to city j by taking action a in city i.
- The negative of the cost can be seen as a reward: $r(i, a, j) = -c(i, j)$

*Only for the exercises solved during the course.

We define:

- New transition probabilities $p(j|i,a)$ such that:

$$\begin{aligned} &p(j|i,a) = \tilde{p}(j|i,a), \text{ if } i \neq s_2 \\ &p(j|s_2,a) = 1, \text{ if } j = s_2 \text{ and } 0 \text{ otherwise} \end{aligned} \tag{3}$$

  this is done because the agent stays at s_2 when it reaches this destination city.

Thus, the goal of the agent is to find a policy π that solves the following optimization problem:

$$\min_{\pi} E_{\pi}\left[\sum_{t=0}^{\infty} r(S_t, A_t, S_{t+1}\middle| S_1 = s_1\right] \tag{4}$$