ULiège
Prof. Damien Ernst
TA : Samy Aittahar

INFO8003-1
Optimal decision making
for complex problems
5th February 2018

Assignment 1

Reinforcement Learning in a Discrete Domain

1 Domain

We describe the domain below :

- State space : $X = \{(x, y) \in \mathbb{N}^2 | x \leq n, y \leq m\}$,
- Action space : $U = \{(1, 0), (-1, 0), (0, 1), (0, -1)\}$,
- Dynamics : $(x_{t+1}, y_{t+1}) = (min(max(x_t + i, 0), n), min(max(y_t + j, 0), m))$ where $(i, j) \in U$,
- Reward signal : $r((x_t, y_t), (i_t, j_t)) = R(g, x_{t+1}, y_{t+1})$,
- Discount factor : $\gamma = 0.99$,
- Time horizon : $T \to +\infty$,

where $g \in \mathbb{R}^{n \times m}$ for any $n, m \in \mathbb{N}$ and $R(g, i, j)$ returns the value of the cell located at (i, j) in g. Note that the reward signal is bounded by $B = \max_{(i,j)} R(g, i, j)$. The domain is deterministic. We provide also a stochastic setting which differs from the deterministic setting by its dynamics, redefined below :

$$(x_{t+1}, y_{t+1}) = \begin{cases} (min(max(x_t + i, 0), n), min(max(y_t + j, 0), m)) & \text{if} \quad w_t \leq 1 - \beta \\ (1, 1) & \text{otherwise,} \end{cases}$$

where $0 \leq w_t \leq 1$ is drawn, following an uniform distribution, at each time step $t = 0, 1, 2...T$ and $0 \leq \beta \leq 1$.

Make sure you (i) consider both settings for each exercise, (ii) rigorously test your implementation at each stage of the assignment and (iii) store any intermediate result to avoid redundant computations.

-3	1	-5	0	19
6	3	8	9	10
5	-8	4	1	-8
6	-9	4	19	-5
-20	-17	-4	-3	9

Figure 1: Domain instance. The number in each cell represents the reward obtained while reaching it after a move. For example, if at time t the agent moves up from the red cell, it receives a reward equal to 5.

2 Implementation of the Domain

Implement the different components of the domain. Test your results by simulating a simple policy (e.g., always go up).

3 Expected Return of a Policy

Implement a routine to estimate J^μ in this domain, where $\mu : X \to U$ is a stationary policy. The implementation needs to handle both deterministic and stochastic settings and should exploit the dynamic programming principle where a sequence of functions J_N for $N = 0, 1, 2...$ is computed using the Bellman Equation. Test your implementation with the following policy : $\mu(x) = (0, 1),\ \forall x \in X$. Display $J^N_\mu(x)$ for each state x.

4 Optimal Policy

Implement a routine which compute $p(x'|x, u)$ and $r(x, u)$ that define the equivalent MDP of the domain, together with γ and T. Afterwards, implement a routine which compute $Q(x, u)$ from these components using the dynamic programming principle. Derive directly μ^* from Q. Display $J^N_{\mu^*}$ for each initial state x.

5 System Identification

Implement a routine which estimates $r(x, u)$ and $p(x'|x, u)$ from a trajectory $h_t = (x_0, u_0, r_0, x_1) ...$ $(x_{t-1}, u_{t-1}, r_{t-1} x_t)$. Display the convergence speed toward p and r. Compute $\widehat{Q}$ with $\widehat{r}(x, u)$ and $\widehat{p}(x'|x, u)$ the MDP estimation of the domain. Derive directly $\widehat{\mu}^*$ from $\widehat{Q}$. Display $J^N_{\widehat{\mu}^*}$, along with $J^N_{\mu^*}$, for each initial state x. Test your implementation on several trajectories. Detail your experiment protocol.

6 Q-LEARNING IN A BATCH SETTING

Implement a routine which computes $\widehat{Q}(x_k, u_k)$ with *Q-learning*. This routine needs to handle both single and multiple trajectories, along with the possibility to do experience replay. Display the convergence speed of $\widehat{Q}$ toward Q. Derive directly $\widehat{\mu}^*$ from $\widehat{Q}$. Display $J^N_{\widehat{\mu}^*}$, along with $J^N_{\mu^*}$, for each initial state x.
Implement an intelligent agent which learns Q with Q-learning and uses an ϵ-greedy policy. The protocol is the following. The agent trains over 100 episodes having 1000 transitions in the domain instance described in Figure 1. An episode always start from state $(3, 3)$. The learning ratio α_k is equal to 0.05, $\forall k$. Test your implementation with different batch sizes for experience replay.

Display $\widehat{Q}$ and $J^N_{\widehat{\mu}^*}$, still along with $J^N_{\mu^*}$, with $\widehat{\mu}^*$ derived from $\widehat{\mu}^*$ under the protocol described above with ϵ equals to 0.05, 0.2, and 0.5. Consider also a set of ϵ's that starts from one of these values and decreases linearly.

Implement another exploration/exploitation policy presented in the scientific literature and test it.

7 DISCOUNT FACTOR

Redo the whole assignment with $\gamma = 0.4$. Compare the differences between the results and discuss them.