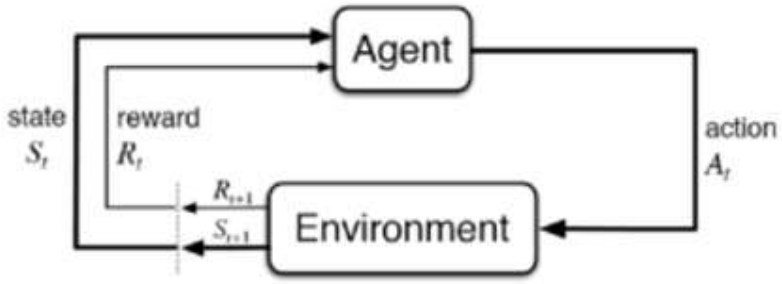
The exercise consists of to complete the next program of Reinforce Learning

Program

The agent can move on a grid. The general target is to reach the final position (0,3) using the minimum steps.



(0, 0)	(0, 1)	(0, 2)	(0, 3) Win (+1)
(1, 0)	(1, 1) Wall	(1, 2)	(1, 3) Lose (-1)
(2, 0) Start	(2, 1)	(2, 2)	(2, 3)

The most of the code is already done, but you have to implement the last section, in main function, to obtain the proper policy of the agent. See attached files.

Here there is some help.

The state (s) is defined as every one of the position of the grid (i , j). That is already defined in the code.

The expected return V(s) is the expected sum of reward from the current position (s). This function is already defined in the code in “evaluate_deterministic_policy(grid, policy)”.

grid is a instance of the environment (game board). See properties in the code.

grid.actions is a dictionary with every state and possible actions according to the state [state | possible action]. For example state (2,0) has Up, and Right.

ACTION_SPACE are all possible actions “a” Up, Down, Right, Left. (U,D,L,R)

policy is dictionary of state and action (s , a)

grid.reward is a dictionary [state | reward]

grid.all_states is set of all state (s) that are in action dictionary or in reward dictionary.

reward is a dictionary [(state , action , next state) | reward] . Tha is the reward associate to go from one state to other state.

transition_probs is a dictionary [(state , action , next state | probability]. All the possible option to go from one state to another has probability 1. The options that are not possible are not included in the dictionary. In this case the value is zero. State is (s), action is denoted by (a), next state is denoted by (s').

Value function Q

$$Q(s,a) = \sum_{s',r} p(s'|s,a) \cdot (r(s,a,s') + GAMMA \cdot V(s'))$$

p(s'|s , a) is the probability of going to the next state (s') given the current state (s), and action (a). That information is in transition_probs

r(s,a,s') is the rewards that is getting when the agent goes from state (s) to next state (s') taking the action (a). This information is in reward dictionary.

Q(s,a) means that the value function is calculate for every state and for every action. So given a state (s) and an action (a), the sum is run for every next state(s').

GAMMA is a constant 0.9

V(s') is the value function calculated in the steps 1.