

## Q Learning A

Consider the grid-world given below and an agent (yellow) moving using these actions: N-North, W-West, E-East, S-South. Rewards are only awarded for reaching one of the terminal states (green and red). Assume discount factor  $\gamma = 1$  for all calculations.

3		-90	80
2			
1	-90	-90	60
	1	2	3

The agent starts from the top left corner and you are given the following episodes from runs of the agent through this grid-world. Each line in an Episode is a tuple containing  $(s_t, a_t, s_{t+1}, r_{t+1})$ , where  $t$  denotes time (iteration) in a training episode.

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5
(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0
(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0
(2,2), S, (2,1), -90	(2,2), S, (2,1), -90	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0	(2,2), N, (2,3), -90
		(3,2), S, (3,1), 60	(3,2), S, (3,1), 60	

Q-learning is an on-line method for learning optimal Q-values in MDP, used in the case of unknown rewards with transitional models. The initial values of the Q function are zero and are updated using the following formulae::

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(\text{trial}_{t,t+1} - Q(s_t, a_t))$$

$$\text{trial}_{t,t+1} = r_{t+1} + \gamma \max_a Q(s_{t+1}, a)$$

and  $\gamma$  is discount factor and  $\alpha$  is learning rate. For following values  $Q$  and upper episodes find out *first episode* and *iteration* ( $t$ ), when value  $Q$  will be non zero. Write it in form E:2, t:3 - in the 2nd episode and 3rd iteration.

$$Q((2,2), E) = \underline{\hspace{2cm}}$$

$$Q((3,2), S) = \underline{\hspace{2cm}}$$

$$Q((1,2), E) = \underline{\hspace{2cm}}$$