Direct Q Evaluation A

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

3		-10	100
2			
1	-30	-30	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, a, s', r).

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), E, (3,2), 0	(2,2), E, (2,1), 0	(2,2), E, (3,2), 0	(2,2), E, (2,3), 0	(2,2), E, (2,1), 0
(3,2), N, (3,3), 0	(2,1), D, (Exit,), -30	(3,2), N, (3,1), 0	(2,3), D, (Exit,), -10	(2,1), D, (Exit,), -30
(3,3), D, (Exit,), 100		(3,1), D, (Exit,), 60		

Fill in the following Q-values obtained using **direct evaluation** from the samples:

$Q((2,2), \text{E}) =$ _____ $Q((1,2), \text{S}) =$ _____ $Q((1,2), \text{E}) =$ _____