Uvažujme kostičkový svět níže. Agent (žlutý) se pohybuje světem pomocí akcí (N-North, W-West, E-East, S-South Reward/odměnu dostane pouze v případě dosažení cílového stavu (zelená a červená políčka). Předpokládejme discount factor $\gamma = 1$.

3		-90	80
2			
1	-90	-70	40
	1	2	3

Agent začíná v levém horním rohu. Vyzkouší několik trénovacích epizod, níže v tabulce. Každý řádek v tabulce trénovací epizody je n-tice $(s_t, a_t, s_{t+1}, r_{t+1})$, kde t indexuje okamžik času (iteraci) v dané učící epizodě. Záznam (1,3), S, (1,2), 0 vlevo nahoře je první iterace $t = 1$ v epizodě 1.

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), -70	(2,2), E, (3,2), 0	(2,2), S, (2,1), -70	(2,2), E, (3,2), 0	(2,2), N, (2,3), -90
	(3,2), S, (3,1), 40		(3,2), N, (3,3), 80	

Q-učení (Q-learning) je on-line metoda pro učení optimálních Q-hodnot v MDP prostředí, kde neznáme odměny (rewards) a přechodové modely. Hodnoty Q funkce průběžně aktualizujeme podle vzorce:

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

kde

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

a γ je faktor zlevnění (discount) a α míra učení (learning rate). Pro následující hodnoty Q a trénovací episody nahoře určete *první episodu* a *iteraci* (t), kdy bude Q hodnota nenulová. Pište ve formě E:2, t:3 - v episodě 2 a iteraci 3.

Q((1,2), E) = _____ Q((2,2), E) = _____ Q((3,2), N) = _____