UNIVERSITÀ DEGLI STUDI DI SALERNO
DIPARTIMENTO DI INFORMATICA

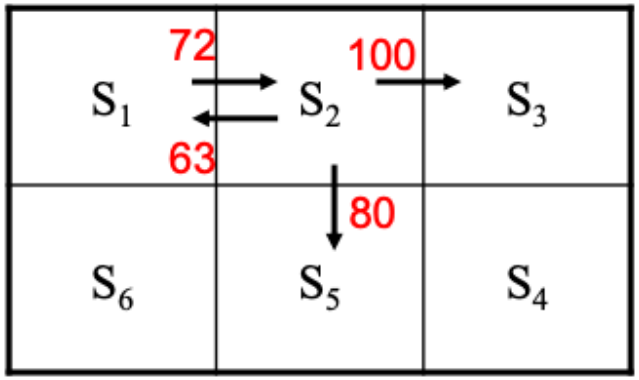
Intelligenza Artificiale

Esercizio Q-Learning

a.a. 2022/2023

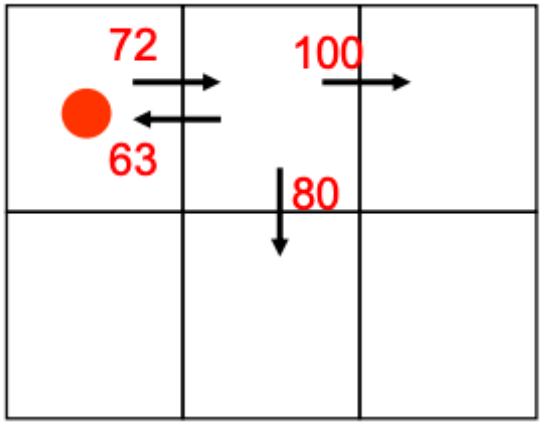
Esempio Q-Learning Update

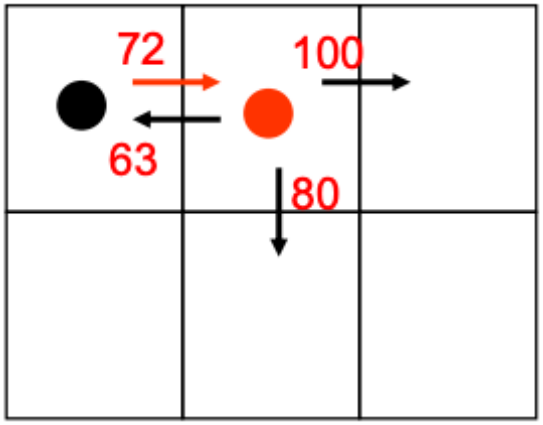
- 6 stati {s1…. s6}
- Azioni: {su, destra, giù, sinistra}
- Reward istananteo = 0
- Inizializzo Q(s,a) – in rosso.



$\gamma = 0.9$

$s_{ini} = s_1$



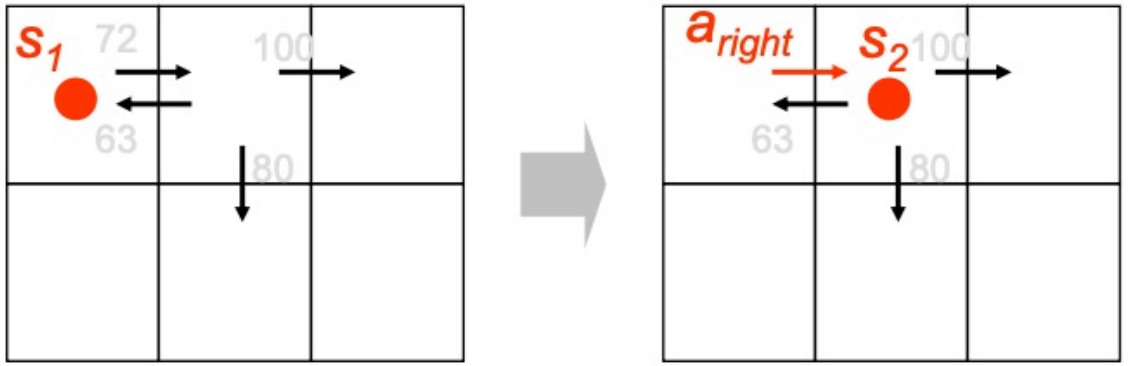


Esempio Q-Learning Update

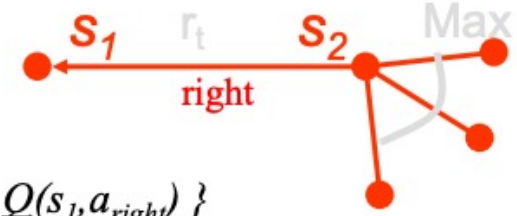
$\gamma = 0.9$

$\alpha = 0.1$

$a(S_2)$ = down



0 reward received in the transition



$$Q(s_1,a_{right}) = Q(s_1,a_{right}) + \alpha\{ r(s_1,a_{right}, s_2) + \gamma \max_{a'} Q(s_2,a') - Q(s_1,a_{right}) \}$$
$$= 72 + \alpha \{0 + 0.9 \max_{a'} \{63, 80, 100\} - Q(s_1,a_{right}) \}$$
$$= 72 + \alpha (90 - 72) = 72 + 1.8 = 73.8$$

Correzione di $Q(s_1,a_{right})$

Correzione dell'azione in s_2 da down a right

La correzione di $Q(s_1,a_{right})$ va a 0 quando

$Q(s_1,a_{right}) = 90$

$Q(s_2,a_{down}) = 80$

$Q(s_2,a_{right}) = 100$

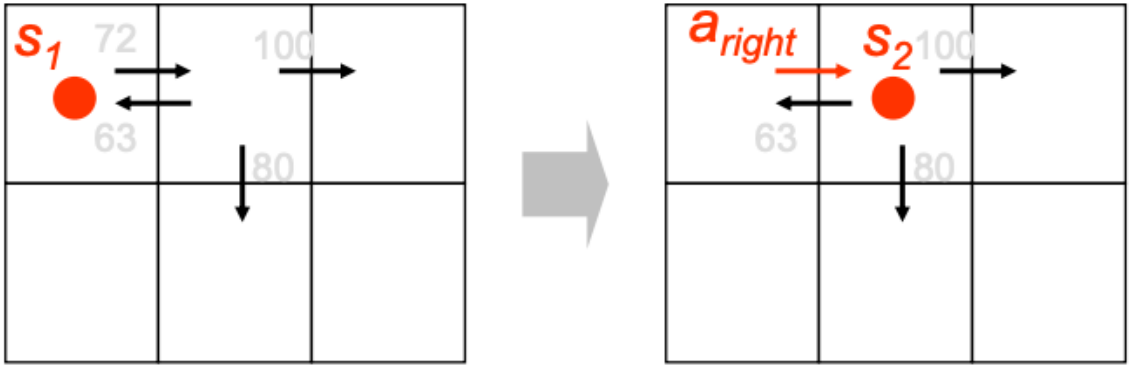
$Q(s_2,a_{left}) = 63$

Esempio Q-Learning Update

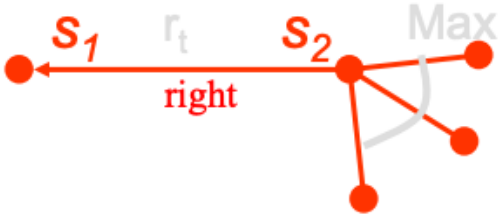
$\gamma = 0.9$

$\alpha = 0.1$

$a(S_2)$ = down



0 reward received in the transition



$Q(s_1, a_{right}) = 72 + \alpha(90 - 72) = 72 + 1.8 = 73.8$ trial 1

$Q(s_1, a_{right}) = 73.8 + \alpha(90 - 73.8) = 73.8 + 1.62 = 75.42$ trial 2

$Q(s_1, a_{right}) = 75.42 + \alpha(90 - 75.42) = 75.42 + 1.458 = 76.878$ trial 3

$Q(s_1, a_{right}) = 76.878 + \alpha(90 - 76.878) = 76.878 + 1.3122 = 78.1902$ trial 4

$Q(s_1, a_{right}) = 78.1902 + \alpha(90 - 78.1902) = 75.42 + 1.458 = 79.37118$ trial 5

$Q(s_1, a_{right}) = 79.37118 + \alpha(90 - 79.37118) = 75.42 + 1.458 = 80.434062$ trial 6

……

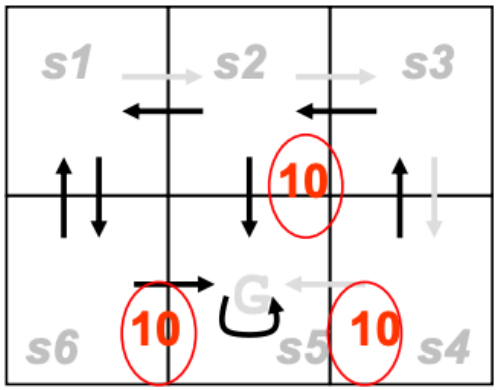
Si ottiene una serie che converge al valore asintotico 90 (asintoticamente)

Esempio 2 - Q-Learning

- Stati: $\{s_1,\ldots s_6\}$
- Azioni: {dx. sx, su, giù}
- **Reward solo in alcune transizioni (in rosso e cerchiato).**
- Stato iniziale: s_1

- Initial selected policy: move clockwise;
- Q(s,a) initially 0;

E.g. videogioco.
In G rimango in G - loop



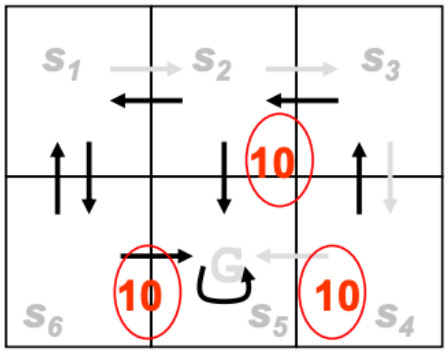
$\alpha = 1$
$\gamma = 0.8.$

Esempio 2 - Q-Learning

- Start at upper left; Initial selected policy: move clockwise; Q(s,a) initially 0; $\gamma = 0.8$. **Reward solo nelle transizioni**.

$$Q_{k+1}^{\pi}(s_1, E) = Q_k^{\pi}(s_1, E) + \alpha \left[r' + \gamma \max_{a'} Q_k^{\pi}(s_2, a') - Q_k^{\pi}(s_1, E) \right]$$

Reward istanteo in rosso e cerchiato



$$Q_{k+1}^{\pi}(s_1, E) = 0 + 1[0 + 0.8x0 - 0] = 0$$

E.g. videogioco.
In G rimango in G - loop

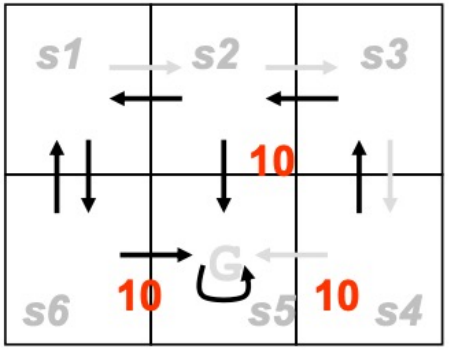
$Q(s_1,East)$	$Q(s_2,East)$	$Q(s_3,South)$	$Q(s_4,West)$
0			

Esempio 2 - Q-Learning

- Start at upper left – move clockwise; table initially 0; $\gamma = 0.8$; $\alpha = 1$

$$Q_{k+1}^{\pi}(s,a) = Q_k^{\pi}(s,a) + \alpha\left[r' + \gamma \max_{a'} Q_k^{\pi}(s',a') - Q_k^{\pi}(s,a)\right]$$

$$Q_{k+1}^{\pi}(s_3,S) = 0 + 1[0 + 0.8x0 - 0] = 0$$



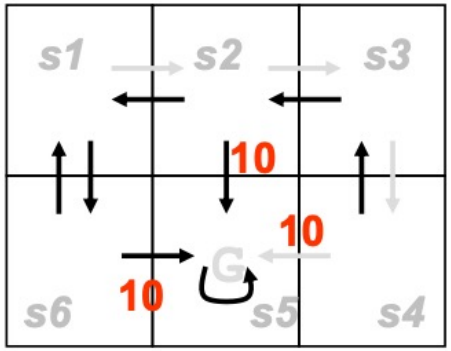
Q(S1,E)	Q(s2,E)	Q(s3,S)	Q(s4,W)
0	0	0	

Esempio 2 - Q-Learning

- Start at upper left – move clockwise; $\gamma = 0.8$

$$Q_{k+1}^{\pi}(s_4, W) = Q_k^{\pi}(s_4, W) + \alpha \left[r' + \gamma \max_{a'} Q_k^{\pi}(s_3, a') - Q_k^{\pi}(s_4, W) \right]$$

$$Q_{k+1}^{\pi}(s_4, W) = 0 + 1[10 + 0.8\ x\ 0 - 0] = 10$$



$Q_k^{\pi}(s_5, .)$ goal

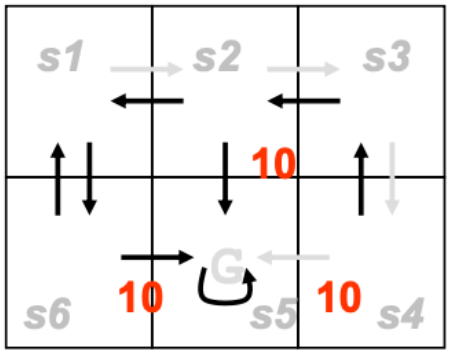
Q(S1,E)	Q(s2,E)	Q(s3,S)	Q(s4,W)
0	0	0	**10**

Esempio 2 - Q-Learning

Start at upper left – move clockwise; $\gamma = 0.8$

$$Q^{\pi}_{k+1}(s_3, S) = Q^{\pi}_{k}(s_3, S) + \alpha \left[r' + \gamma \max_{a'} Q^{\pi}_{k}(s_4, a') - Q^{\pi}_{k}(s_3, S) \right]$$

$$Q^{\pi}_{k+1}(s_3, S) = 0 + 1[0 + 0.8\,\{max, 10, 0\} - 0] = 8$$



$Q^{\pi}_{k}(s_4, W)$

$Q^{\pi}_{k}(s_4, N)$

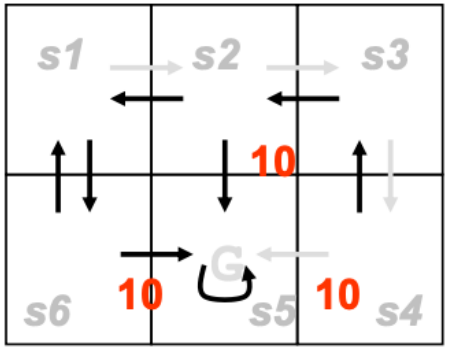
Q(S1,E)	Q(s2,E)	Q(s3,S)	Q(s4,W)
0	0	0	10
0	0	**8**	

Esempio 2 - Q-Learning

- Start at upper left – move clockwise; $\gamma = 0.8$

$$Q_{k+1}^{\pi}(s_4, W) = Q_k^{\pi}(s_4, W) + \alpha \left[r' + \gamma \max_{a'} Q_k^{\pi}(s_3, a') - Q_k^{\pi}(s_4, W) \right]$$

$$Q_{k+1}^{\pi}(s_4, W) = 10 + 1[10 + 0.8\ x\ 0 - 10] = 10$$



$Q_k^{\pi}(s_5, .)$ goal

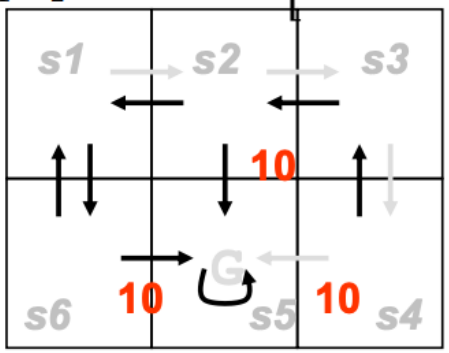
Q(S1,E)	Q(s2,E)	Q(s3,S)	Q(s4,W)
0	0	0	10
0	0	8	10

Esempio 2 - Q-Learning

- Start at upper left – move clockwise; $\gamma = 0.8$

$$Q_{k+1}^{\pi}(s_2, E) = Q_k^{\pi}(s_2, E) + \alpha \left[r' + \gamma \max_{a'} Q_k^{\pi}(s_3, a') - Q_k^{\pi}(s_2, E) \right]$$

$$Q_{k+1}^{\pi}(s_2, E) = 0 + 1 \left[0 + 0.8 \; x \max_{a'} \{8 \, , 0\} - 0 \right] = 6.4$$



$Q_k^{\pi}(s_3, S)$ $Q_k^{\pi}(s_3, W)$

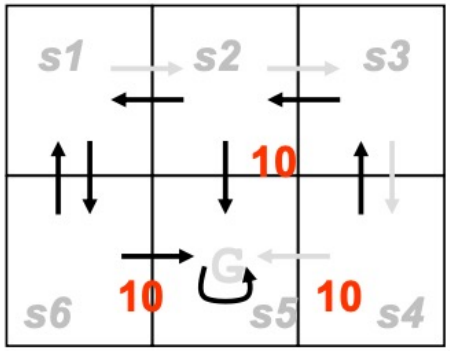
Q(S1,E)	Q(s2,E)	Q(s3,S)	Q(s4,W)
0	0	0	10
0	0	8	10
0	**6.4**		

Esempio 2 - Q-Learning

- Start at upper left – move clockwise; $\gamma = 0.8$

$$Q^{\pi}_{k+1}(s_3, S) = Q^{\pi}_{k}(s_3, S) + \alpha \left[r' + \gamma \max_{a'} Q^{\pi}_{k}(s_4, a') - Q^{\pi}_{k}(s_3, S) \right]$$

$$Q^{\pi}_{k+1}(s_3, S) = 0 + 1[0 + 0.8\ \{max, 10, 0\} - 0] = 8$$



$Q^{\pi}_{k}(s_4, W)$

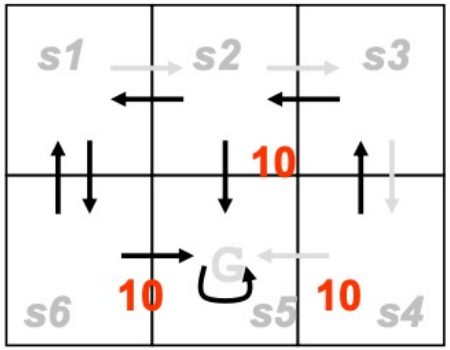
$Q^{\pi}_{k}(s_4, N)$

Q(S1,E)	Q(s2,E)	Q(s3,S)	Q(s4,W)
0	0	0	10
0	0	8	10
0	**6.4**	8	10

Esempio 2 - Q-Learning

$$Q^{\pi}_{k+1}(s_1, E) = Q^{\pi}_{k}(s_1, E) + \alpha \left[r' + \gamma \max_{a'} Q^{\pi}_{k}(s_2, a') - Q^{\pi}_{k}(s_1, E) \right]$$

$$Q^{\pi}_{k+1}(s_1, E) = 0 + 1 \left[0 + 0.8 \; x \max_{a'} \{6.4, 0, 0\} - 0 \right] = 5.12$$



$Q^{\pi}_{k}(s_2, W)$

$Q^{\pi}_{k}(s_2, E)$

$Q^{\pi}_{k}(s_2, S)$

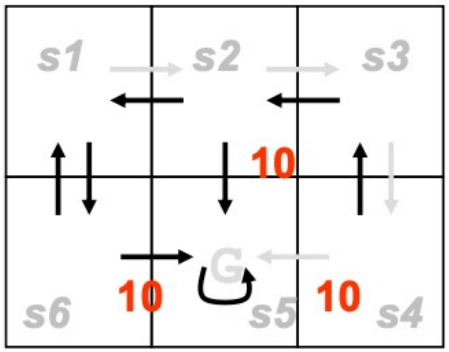
Q(S1,E)	Q(s2,E)	Q(s3,S)	Q(s4,W)
0	0	0	10
0	0	8	10
0	6.4	8	10
5.12	6.4	8	10

Esempio 2 - Q-Learning

- Start at upper left – move clockwise; $\gamma = 0.8$; $\alpha = 1$

$$Q^{\pi}_{k+1}(s_2, S) = Q^{\pi}_{k}(s_2, S) + \alpha \left[r' + \gamma \max_{a'} Q^{\pi}_{k}(s_5, a') - Q^{\pi}_{k}(s_2, S) \right]$$

$$Q^{\pi}_{k+1}(s_2, S) = 0 + 1[10 + 0.8\ x\ 0 - 0] = 10$$



$Q^{\pi}_{k}(s_5, .)$

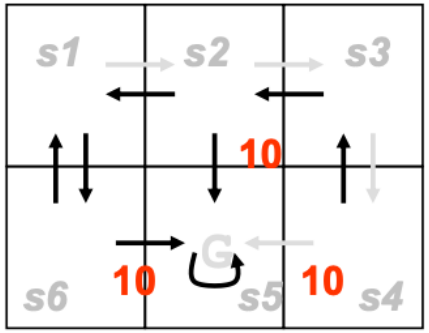
Mossa ε-greedy in s_2 (invece che a = E, scelgo a = S, cambio azione):
calcolo $Q(s_2, S) = r + \gamma \max_{a'} \{Q(s5,a')\} = 10 + 0.8 \times 0 = \mathbf{10}$

Q(S1,E)	Q(s2,E)	Q(s2,S)	Q(s3,S)	Q(s4,W)
0	0	0	0	10
0	0	0	8	10
0	6.4	0	8	10
5.12	6.4	10	8	10

Esempio 2 - Q-Learning

$$Q^{\pi}_{k+1}(s_1, E) = Q^{\pi}_{k}(s_1, E) + \alpha \left[r' + \gamma \max_{a'} Q^{\pi}_{k}(s_2, a') - Q^{\pi}_{k}(s_1, E) \right]$$

$$Q^{\pi}_{k+1}(s_1, E) = 5.12 + 1 \left[0 + 0.8 \; x \max_{a'} \{6.4, 10, 0\} - 5.12 \right] = 8$$



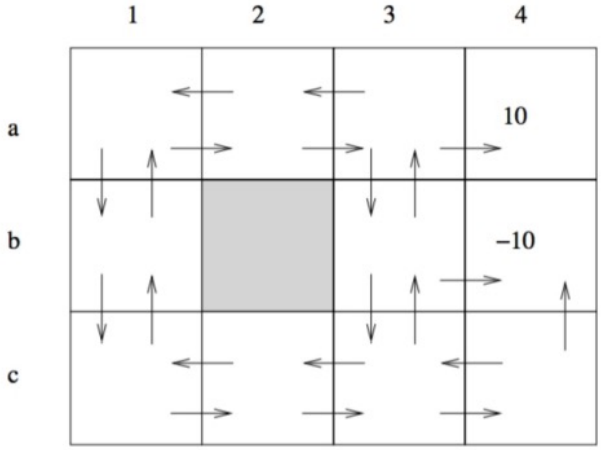
$Q^{\pi}_{k}(s_2, W)$

$Q^{\pi}_{k}(s_2, E)$

$Q^{\pi}_{k}(s_2, S)$

Q(S1,E)	Q(s2,E)	Q(s2,S)	Q(s3,S)	Q(s4,W)
0	0	0	0	10
0	0	0	8	10
0	6.4	0	8	10
8	6.4	10	8	10

Esercizio



- Simulare l'algoritmo Q-learning per un robot che cammina nell'ambiente in figura (b2 è un muro, entrare in b4 da una penalità di -10, entrare in a4 fornisce una ricompensa di 10).
- Indicare i Q-value dopo i seguenti episodi, usando l'aggiornamento all'indietro (dopo essere arrivati allo stato obiettivo, i Q-value aggiornano in ordine inverso dal goal all'inizio, e γ =0.9)

1. a1, a2, a3, b3, b4
2. c2, c1, b1, a1, a2, a3, a4
3. c4, c3, b3, a3, a4

- Indicare la policy che esegue l'azione che ha il Q-value più alto. E' ottimale?