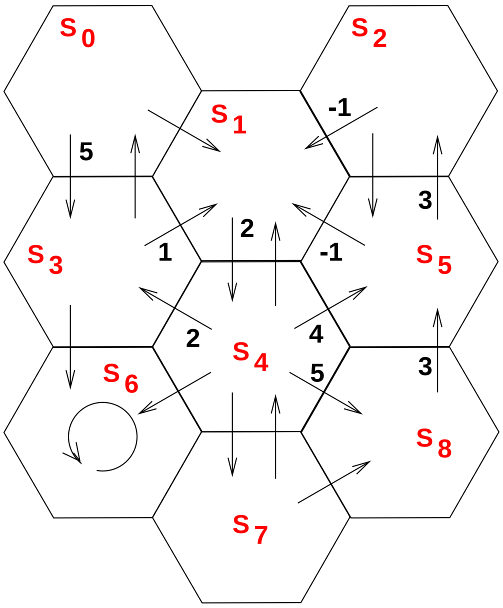
Esercizi per la Seconda Parte

Anno Accademico 00/01

Esercizio 3

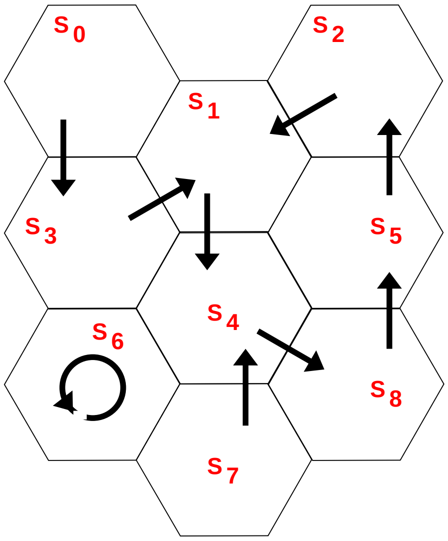
a1) Si consideri il seguente ambiente con le corrispondenti azioni e ricompense



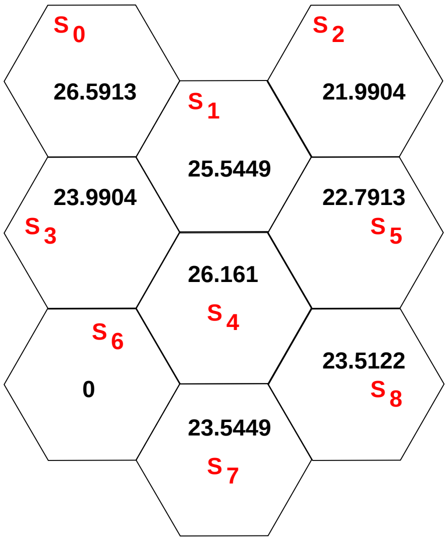
Calcolare i valori di V^* per ogni stato e $Q(s, a)$ per ogni coppia (s, a) avendo posto $\gamma = 0.9$.

Risposta:

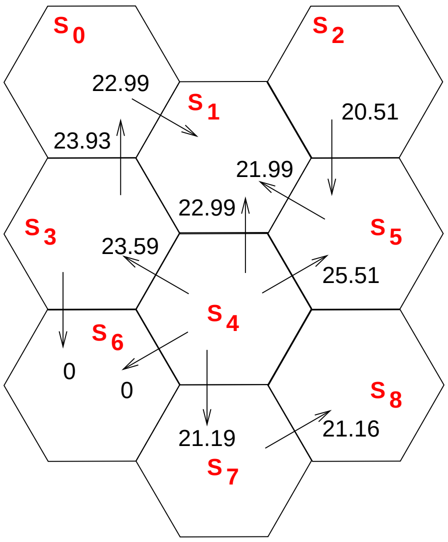
POLITICA OTTIMA



V*



Q(s,a)



Le azioni della politica ottima hanno valore Q(s,a) uguale al valore V* dello stato di partenza. Si riportano i valori delle altre azioni.

a2) Scoprire il valore minimo di γ per il quale la politica ottima definita in a1) rimane valida, cioè quel valore γ^{min} tale che per $\gamma < \gamma^{min}$ la politica ottima in a1) cambia.

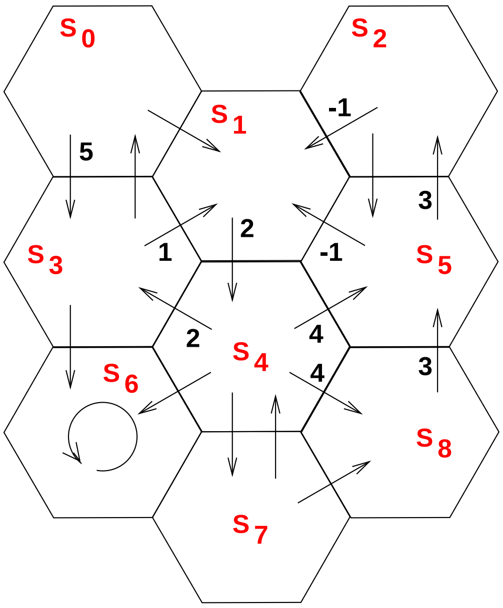
Risposta:

Poiché si considerano valori di $\gamma < 0.9$ il "percorso lungo" $s_4 \to s_8 \to s_5 \to s_2 \to s_1 \to s_4$ della politica ottima viene "spezzato" dal ciclo $s_5 \to s_2 \to s_5$, che diventa più conveniente con il decrescere di γ. Quindi si deve determinare il valore di $\gamma < 0.9$ per cui:

$$\frac{5 + 3\gamma + 3\gamma^2 - \gamma^3 + 2\gamma^4}{1 - \gamma^5} < 5 + 3\gamma + \frac{3\gamma^2}{1 - \gamma^2}$$

Si noti che il percorso $s_4 \to s_3 \to s_0 \to s_3$ non risulta essere migliore del percorso $s_4 \to s_8 \to s_5 \to s_2 \to s_1 \to s_4$ per valori di $\gamma < 0.9$.

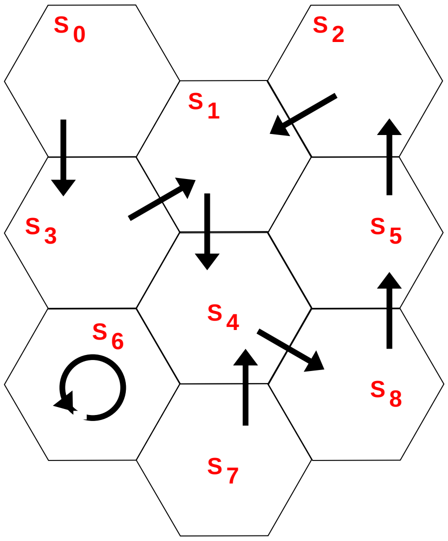
b2) Si consideri il seguente ambiente con le corrispondenti azioni e ricompense (le frecce senza numero aasociato indicano una ricompensa immediata di 0) Calcolare i valori



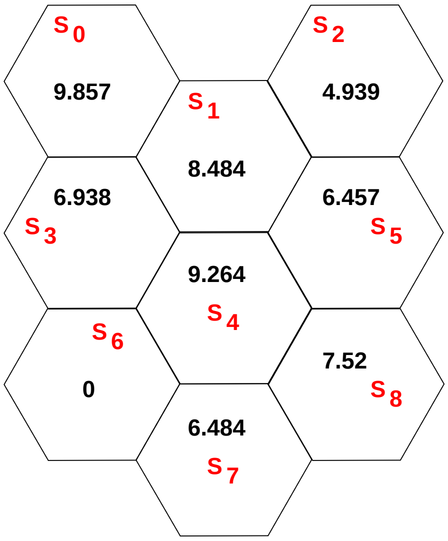
di V^* per ogni stato e $Q(s, a)$ per ogni coppia (s, a) avendo posto $\gamma = 0.7$.

Risposta:

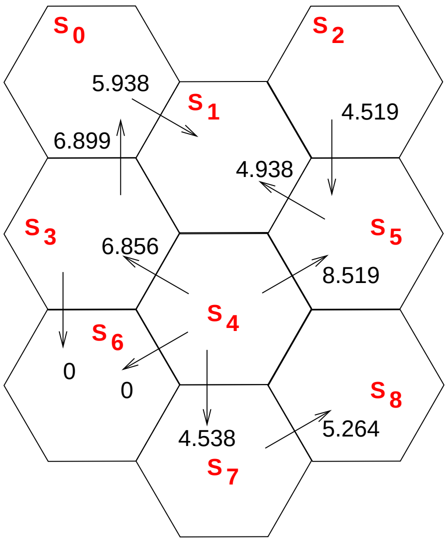
POLITICA OTTIMA



V*

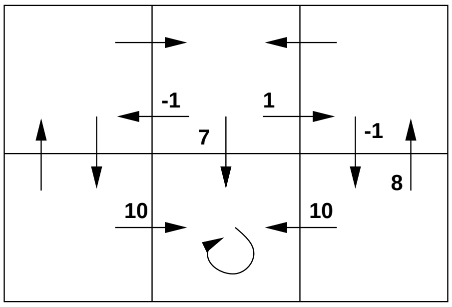


Q(s,a)



Le azioni della politica ottima hanno valore Q(s,a) uguale al valore V* dello stato di partenza. Si riportano i valori delle altre azioni.

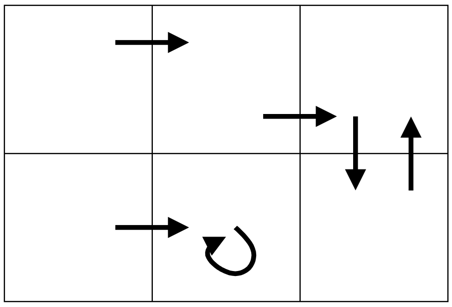
c2) Si consideri il seguente ambiente con le corrispondenti azioni e ricompense. Calcolare



i valori di V^* per ogni stato e $Q(s,a)$ per ogni coppia (s,a) avendo posto $\gamma = 0.8$;

Risposta:

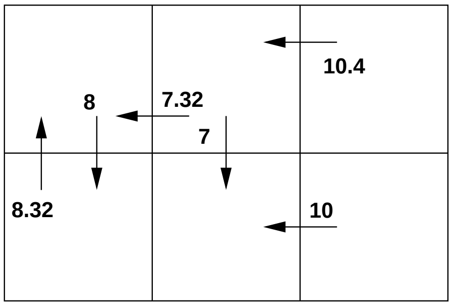
POLITICA OTTIMA



V*

10.4	**13**	**15**
10	**0**	**20**

Q(s,a)



Le azioni della politica ottima hanno valore Q(s,a) uguale al valore V* dello stato di partenza. Si riportano i valori delle altre azioni.

c3) Se con $(0,0)$ si indica lo stato in basso a sinistra, si mostri l'evoluzione dell'algoritmo Q-learning per la seguente sequenza (con reset dopo aver raggiunto lo stato $(1,0)$) di esplorazione:

$$(0,0),(0,1),(0,0),(1,0),reset,(2,1),(1,1),(2,1),(2,0),(1,0)$$

Risposta:

	→	←	↑	↓	↻
(0,0)		**no**		**no**	**no**
(1,0)	**no**	**no**	**no**	**no**	
(2,0)	**no**			**no**	**no**
(0,1)		**no**	**no**		**no**
(1,1)			**no**		**no**
(2,1)	**no**		**no**		**no**

Le caselle vuote contengono 0

s = (0,0) a = ↑

	→	←	↑	↓	↻
(0,0)		**no**		**no**	**no**
(1,0)	**no**	**no**	**no**	**no**	
(2,0)	**no**			**no**	**no**
(0,1)		**no**	**no**		**no**
(1,1)			**no**		**no**
(2,1)	**no**		**no**		**no**

Le caselle vuote contengono 0

s = (0,1) a = ↓

	→	←	↑	↓	↻
(0,0)	10	**no**		**no**	**no**
(1,0)	**no**	**no**	**no**	**no**	
(2,0)	**no**			**no**	**no**
(0,1)		**no**	**no**		**no**
(1,1)			**no**		**no**
(2,1)	**no**		**no**		**no**

Le caselle vuote contengono 0

s = (0,0) a = →

	→	←	↑	↓	↻
(0,0)	10	**no**		**no**	**no**
(1,0)	**no**	**no**	**no**	**no**	
(2,0)	**no**			**no**	**no**
(0,1)		**no**	**no**		**no**
(1,1)			**no**		**no**
(2,1)	**no**		**no**		**no**

Le caselle vuote contengono 0

s = (2,1) a = ←

	→	←	↑	↓	↻
(0,0)	10	**no**		**no**	**no**
(1,0)	**no**	**no**	**no**	**no**	
(2,0)	**no**			**no**	**no**
(0,1)		**no**	**no**		**no**
(1,1)	1		**no**		**no**
(2,1)	**no**		**no**		**no**

Le caselle vuote contengono 0

s = (1,1) a = →

	→	←	↑	↓	↻
(0,0)	10	**no**		**no**	**no**
(1,0)	**no**	**no**	**no**	**no**	
(2,0)	**no**			**no**	**no**
(0,1)		**no**	**no**		**no**
(1,1)	1		**no**		**no**
(2,1)	**no**		**no**	−1	**no**

Le caselle vuote contengono 0

s = (2,1) a = ↓

	→	←	↑	↓	↻
(0,0)	10	**no**		**no**	**no**
(1,0)	**no**	**no**	**no**	**no**	
(2,0)	**no**	10		**no**	**no**
(0,1)		**no**	**no**		**no**
(1,1)	1		**no**		**no**
(2,1)	**no**		**no**	−1	**no**

Le caselle vuote contengono 0

s = (2,0) a = ←

(0,1)	(1,1)	(2,1)
(0,0)	(1,0)	(2,0)