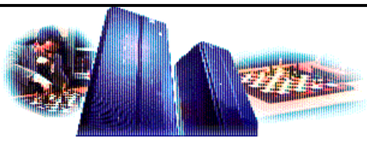
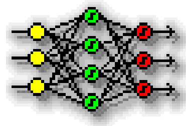




Sistemi Intelligenti Reinforcement Learning: Temporal Differences

Alberto Borghese

Università degli Studi di Milano
Laboratorio di Sistemi Intelligenti Applicati (AIS-La
Dipartimento di Informatica
alberto.borghese@unimi.it

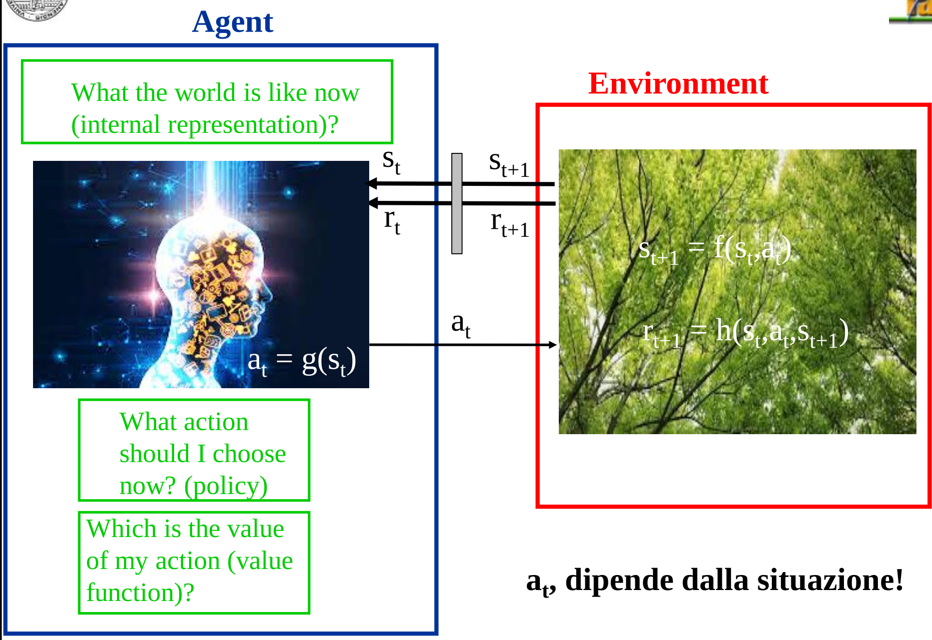
Barto and Sutton, Capitoli 3 e 6

Sommario

Le equazioni di Bellman

Differenze temporali

The RL updated picture



a_t, dipende dalla situazione!

Meccanismo di apprendimento nel RL

Inizializzazione: se l’agente non agisce sull’ambiente non succede nulla. Occorre specifiare una policy iniziale.

Ciclo dell’agente (le tre fasi sono sequenziali):

1) Implemento una policy ($\pi(s,a)$)
2) **Aggiorno la Value function ($Q^{\pi}(s,a)$)**
3) Aggiorno la policy.

http:\\borghese.di.unimi.it\

Esempio: AIBO search

Azioni:
1) Rimanere fermo e aspettare che qualcuno getti nel cestino una lattina vuota.
2) Muoversi attivamente in cerca di lattine.
3) Tornare alla sua base (recharge station) e ricaricarsi.

Stato:
1) Alto livello di energia.
2) Basso livello di energia.

Azioni ammissibili (policy):
a(s = high) = {Search, Wait}
a(s = low) = {Search, Wait, Recharge}

Goal: collezionare il maggior numero di lattine.

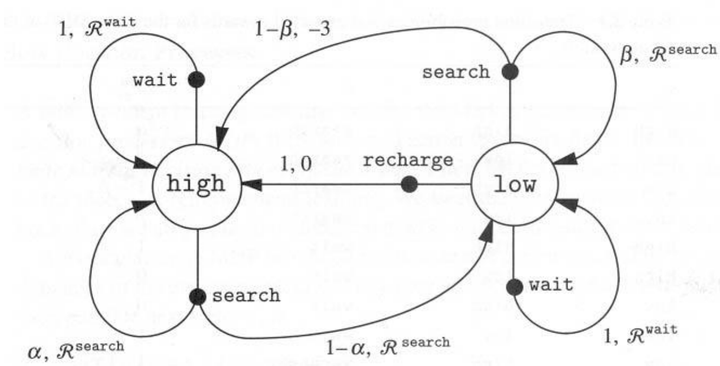
http:\\borghese.di.unimi.it\

Esempio di calcolo della Value function

Policy deterministica
a(high) = wait
a(low) = search

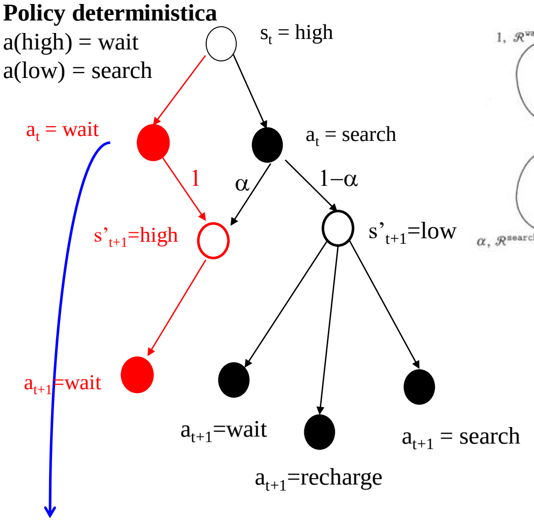
Value function
Q(high, search) = ?
Q(low, search) = ?

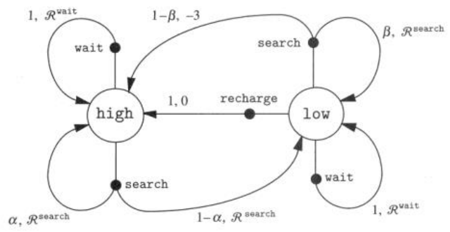
α=Pr($s_{t+1} = High | s_t = High, a_t = Search$) =0.4
β=Pr($s_{t+1} = Low | s_t = Low, a_t = Search$)=0.1,
γ=0.8, R^{search}=3, R^{wait}=1 R^{dead} = -3, R^{auto} = 0



Analisi ad un passo dal tempo t

Policy deterministica
a(high) = wait
a(low) = search





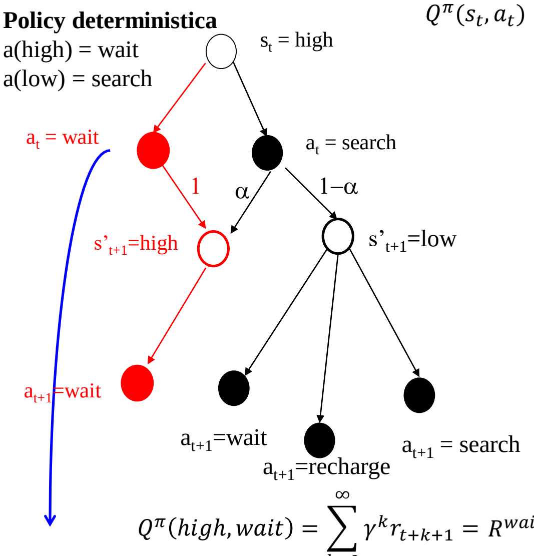
$$Q^{\pi}(s_t, a_t) = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

$$Q^{\pi}(s_t, a_t) = E_{\pi}\{R_t | s_t = s, a_t = a\}$$

Analisi ad un passo dal tempo t

Policy deterministica
a(high) = wait
a(low) = search

$$Q^{\pi}(s_t, a_t) = E_{\pi}\{R_t | s_t = s, a_t = a\}$$



$$Q^{\pi}(high, wait) = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} =$$

$$R^{wait} + \sum_{k=1}^{\infty} \gamma^k r_{t+k+1} =$$

$$R^{wait} + \gamma \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k+1} =$$

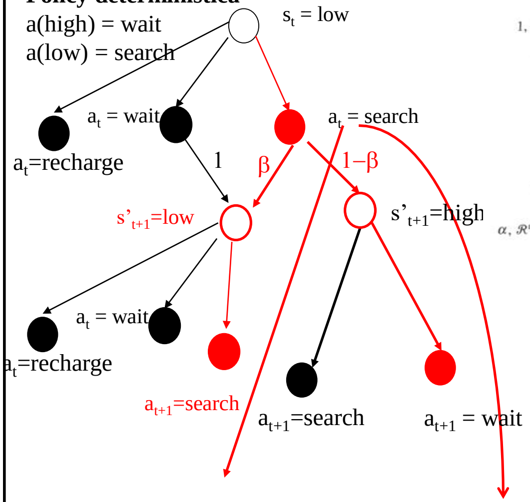
$$R^{wait} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2}$$

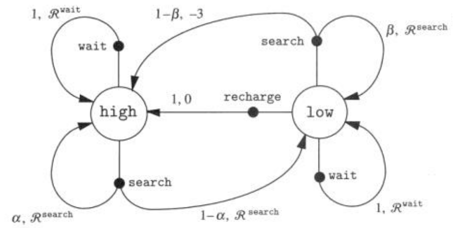
$$Q^{\pi}(high, wait) = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} = R^{wait} + \gamma Q^{\pi}(high, wait)$$

$Q^{\pi}(h,w) = [1 + 0.8\, Q^{\pi}(h,w)]$

Analisi ad un passo dal tempo t

Policy deterministica
a(high) = wait
a(low) = search





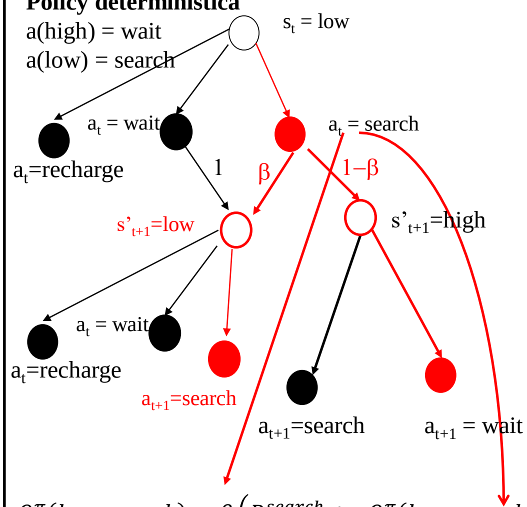
$$Q^{\pi}(s_t, a_t) = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

2 cammini possibili!!

$$Q^{\pi}(s_t, a_t) = E_{\pi}\{R_t | s_t = s, a_t = a\}$$

Analisi ad un passo dal tempo t

Policy deterministica
a(high) = wait
a(low) = search

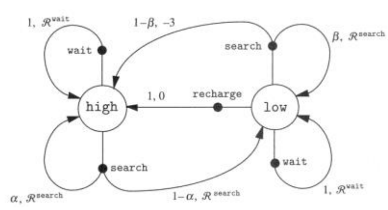


$$Q^{\pi}(low, search) = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} =$$

2 cammini possibili!!

1) $R^{search} + \gamma Q^{\pi}(low, search)$

2) $R^{dead} + \gamma Q^{\pi}(high, wait)$



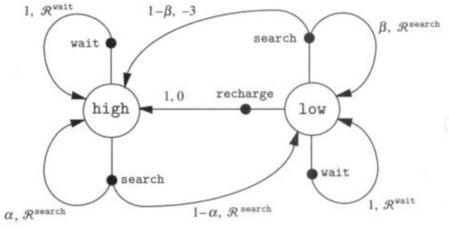
$$Q^{\pi}(low, search) = \beta\left(R^{search} + \gamma Q^{\pi}(low, search)\right) + \quad (1-\beta)\left(R^{dead} + \gamma Q^{\pi}(high, wait)\right)$$

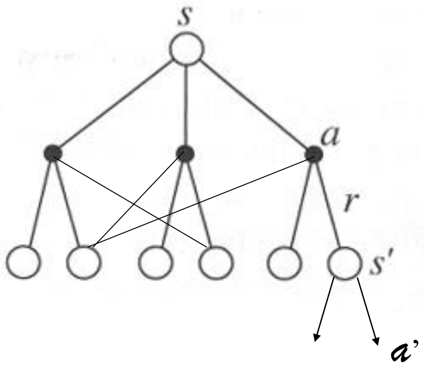
Q(l,s) = 0.1x[3+0.8xQ(l,s)]+0.9x[-3+0.8 Q(h,w)]

Contiene la probabilità di ricevere un reward γQ(s',a), condizionata a s_{t+1} = s'!

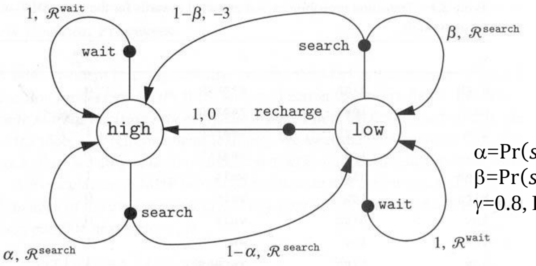
Valutazione policy stocastica

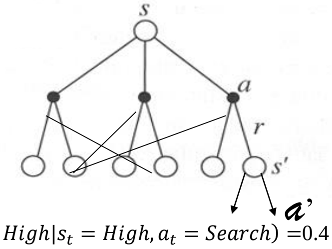
Nel valutare Q(s,a) dobbiamo valutare tutti i cammini che partono da ogni s’.





Policy stocastica





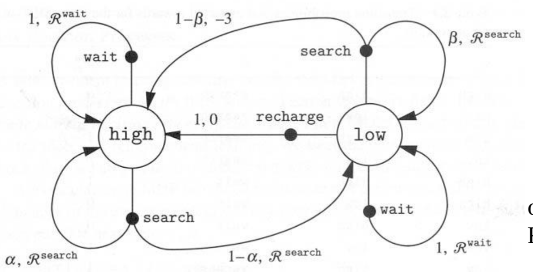
α=$\Pr(s_{t+1} = High|s_t = High, a_t = Search)$ =0.4
β=$\Pr(s_{t+1} = Low|s_t = Low, a_t = Search)$=0.1,
γ=0.8, R^{search}=3, R^{wait}=1 R^{dead} = -3, R^{auto} = 0

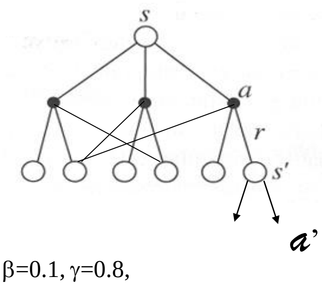
Q(high,wait) = 1 x {R^{wait}+γ [Pr(a’=search|high) **Q(high,search)**+ Pr(a’=wait|high) **Q(high,wait)**]}
Q(high,wait) = 1x {1+0.8[Pr(a’=search|high) **Q(high,search)** + Pr(a’=wait|high) **Q(high,wait)**]}

Q(high,search) = $\Pr(s_{t+1} = High|s_t = High, a_t = Search)$ x
{R^{search}+γ [(Pr(a’=search|high) **Q(high,search)** + Pr(a’=wait|high) **Q(high,wait)**]} +
(1- $\Pr(s_{t+1} = High|s_t = High, a_t = Search)$) x
{R^{search}+γ [Pr(a’=search|low)**Q(low,search)**+Pr(a’=wait|low)**Q(low,wait)**+Pr(a’=recharge|low)**Q(low,rech)]}**

Q(high,search) = 0.4x{3+0.8[(Pr(a’=search|high)**Q(high,search)**+Pr(a’=wait|high) **Q(high,wait)**]} +
0.6x{3+0.8[Pr(a’=search|low)**Q(low,search)**+Pr(a’=wait|low)**Q(low,wait)**
+Pr(a’=recharge|low)**Q(low,rech)]}**

Policy stocastica





α=0.4, β=0.1, γ=0.8,
R^{search}=3, R^{wait}=1 R^{dead} = -3, R^{auto} = 0

Q(low,wait) = 1 x {R^{wait}+γ [Pr(a'=search) **Q(low,search)**+ Pr(a'=wait) **Q(low,wait)** + Pr(a'=recharge) **Q(low,reharge)**]}

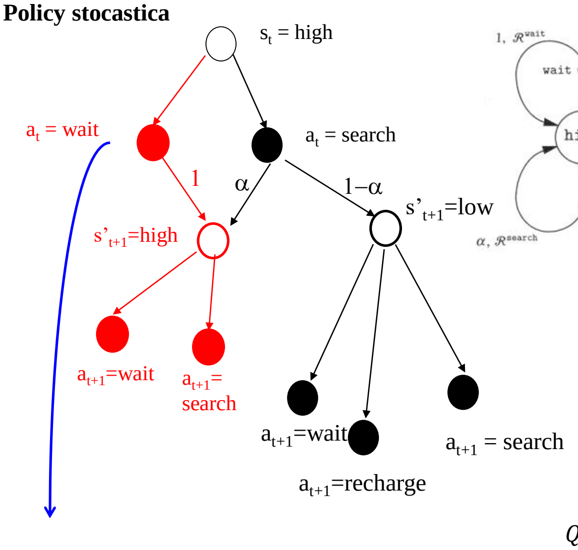
Q(low,search) = β x {R^{search}+γ [(Pr(a'=search) **Q(high,search)**+ Pr(a'=wait) **Q(high,wait)+** Pr(a'=recharge) **Q(low,reharge)**]} +
(1-β) x {R^{dead}+γ [Pr(a'=search) **Q(high,search)**+ Pr(a'=wait) **Q(high,wait)]}**

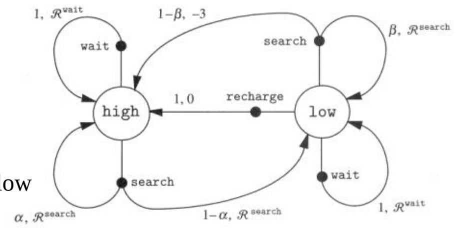
Q(low,recharge) = 1 x {R^{auto}+γ [(Pr(a'=search) **Q(high,search)**+ Pr(a'=wait) **Q(high,wait)**]}

5 equazioni in 5 incognite

Analisi ad un passo dal tempo t

Policy stocastica



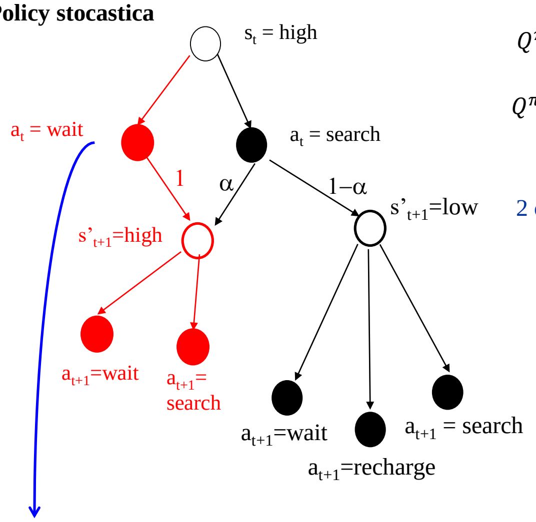


$$Q^{\pi}(s_t, a_t) = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

$$Q^{\pi}(s_t, a_t) = E_{\pi}\{R_t | s_t = s, a_t = a\}$$

Analisi ad un passo dal tempo t

Policy stocastica



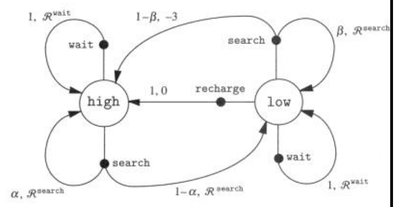
$$Q^{\pi}(s_t, a_t) = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

$$Q^{\pi}(s_t, a_t) = E_{\pi}\{R_t | s_t = s, a_t = a\}$$

2 cammini possibili!!

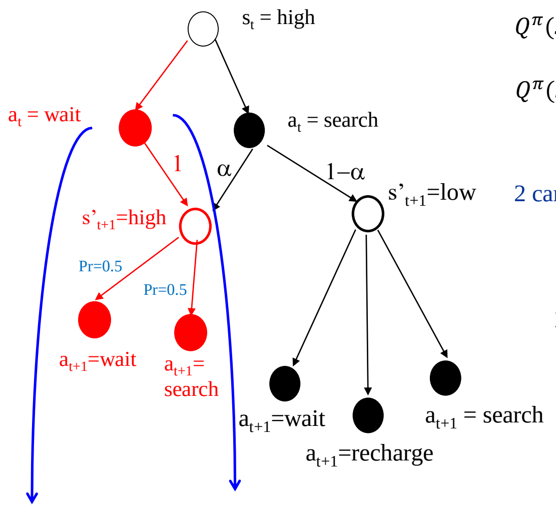
1) $R^{wait} + \gamma Q^{\pi}(high, wait)$

2) $R^{wait} + \gamma Q^{\pi}(high, search)$



Analisi ad un passo dal tempo t

Policy stocastica (uniforme)



$$Q^{\pi}(s_t, a_t) = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

$$Q^{\pi}(s_t, a_t) = E_{\pi}\{R_t | s_t = s, a_t = a\}$$

2 cammini possibili!!

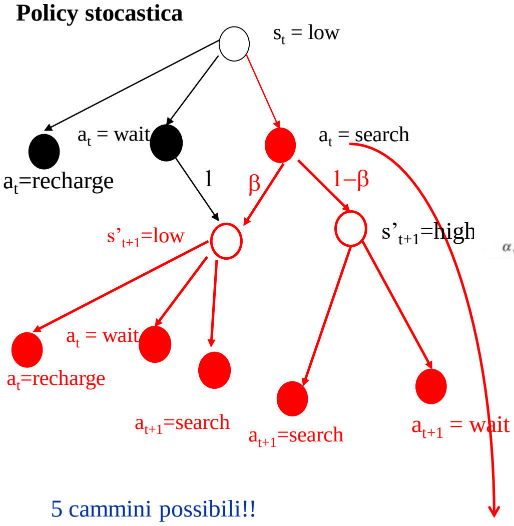
1) $R^{wait} + \gamma Q^{\pi}(high, wait)$

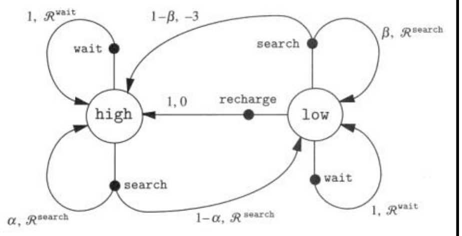
2) $R^{wait} + \gamma Q^{\pi}(high, search)$

$$Q^{\pi}(high, wait) = R^{wait} + 0.5\, \gamma\, Q^{\pi}(high, wait) + 0.5 \gamma Q^{\pi}(high, search)$$

Analisi ad un passo dal tempo t

Policy stocastica





$$Q^{\pi}(s_t, a_t) = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

$$Q^{\pi}(s_t, a_t) = E_{\pi}\{R_t | s_t = s, a_t = a\}$$

5 cammini possibili!!

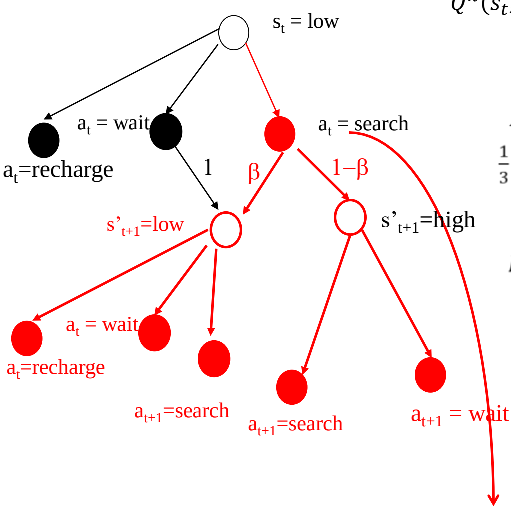
$$Q^{\pi}(low, search) = E_{\pi}\{R_t | s_t = low, a_t = search\}$$

Analisi ad un passo dal tempo t

Policy stocastica (equiprobabile)

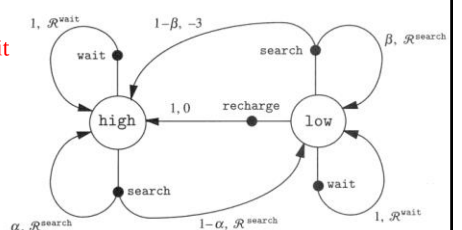
$$Q^{\pi}(s_t, a_t) = E_{\pi}\{R_t | s_t = s, a_t = a\}$$

5 cammini possibili!!



$$A)\ R^{search} + \gamma[\frac{1}{3}Q^{\pi}(low, search) + \frac{1}{3}Q^{\pi}(low, wait) + \frac{1}{3}Q^{\pi}(low, recharge)]$$

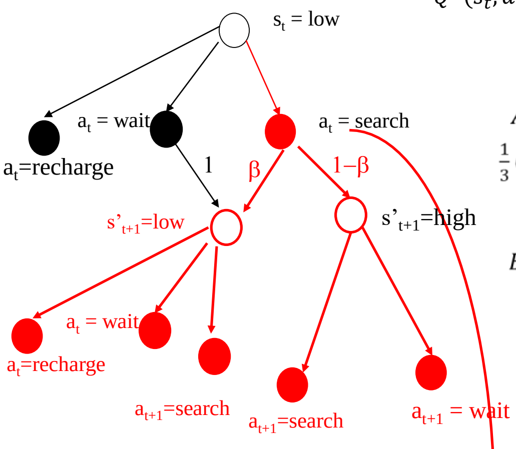
$$B)\ R^{dead} + \gamma[\frac{1}{2}Q^{\pi}(high, search) + \frac{1}{2}Q^{\pi}(high, wait)]$$



Analisi ad un passo dal tempo t

Policy stocastica (equiprobabile)

$$Q^\pi(s_t, a_t) = E_\pi\{R_t | s_t = s, a_t = a\}$$



5 cammini possibili!!

$$A)\ R^{search} + \gamma[\frac{1}{3}Q^\pi(low, search) + \frac{1}{3}Q^\pi(low, wait) + \frac{1}{3}Q^\pi(low, recharge)]$$

$$B)\ R^{dead} + \gamma[\frac{1}{2}Q^\pi(high, search) + \frac{1}{2}Q^\pi(high, wait)]$$

$$Q^\pi(low, search) = \beta[R^{search} + \gamma(\frac{1}{3}Q^\pi(low, search) + \frac{1}{3}Q^\pi(low, wait) + \frac{1}{3}Q^\pi(low, recharge))] +$$
$$(1-\beta)\ [R^{dead} + \gamma(\frac{1}{2}Q^\pi(high, search) + \frac{1}{2}Q^\pi(high, wait))]$$

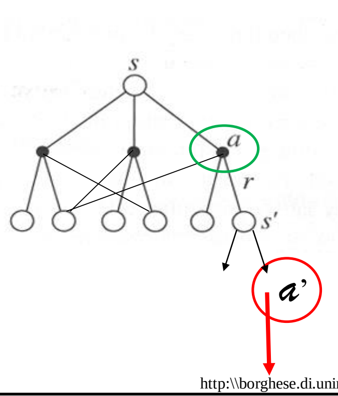
5 equazioni in 5 incognite

Calcolo ricorsivo della Value function

$$\boxed{Q^\pi(s_t, a_t) = E_\pi\{R_t | s_t = s, a_t = a\} = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}}$$

$$\boxed{Q^\pi(s_{t+1}, a_{t+1}) = E_\pi\{R_t | s_{t+1} = s', a_{t+1} = a'\}}$$

Relazione tra $Q^\pi(s,a)$ e $Q^\pi(s',a')$?



Calcolo ricorsivo della Value function

$$\boxed{Q^{\pi}(s_t, a_t) = E_{\pi}\{R_t | s_t = s, a_t = a\} = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}}$$

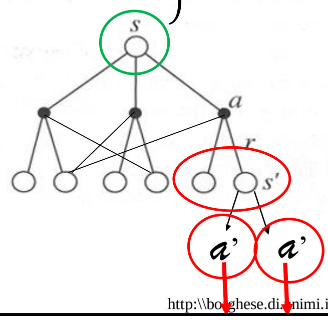
Isolo il reward ad un passo nella serie dei reward.

$$Q^{\pi}(s_t, a_t) = E_{\pi}\{\gamma^0 r_{t+1} + \sum_{k=1}^{\infty} \gamma^k r_{t+k+1} \,| s_t = s, a_t = a\} \Rightarrow$$

$$Q^{\pi}(s_t, a_t) = E_{\pi}\left\{\gamma^0 r_{t+1} + \sum_{k=0}^{\infty} \gamma^{k+1} r_{t+k+2} \,| s_t = s, a_t = a\right\}$$

Io termine
(a un passo)

IIo termine
(passi futuri)

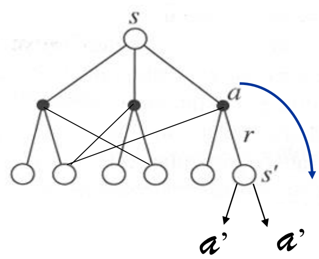


$Q^{\pi}(s,a)$: primo termine

$$\mathrm{P}_{s \to s'|a} \triangleq \Pr(s_{t+1} = s' | s_t = s, a_t = a)$$

$$E_{\pi}\{r_{t+1} | s_t = s, a_t = a\} = \sum_{s'} \mathrm{P}_{s \to s'|a} R_{s,s',a}$$

Per ogni stato-azione devo valutare:
- Più stati prossimi
- Reward stocastici nella transizione ad un passo

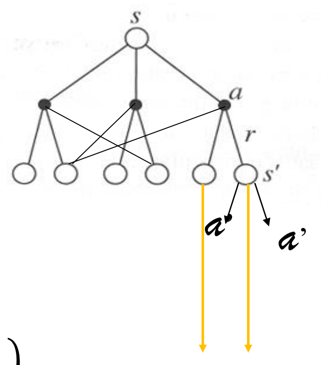


Visione Statistica: Probabilità di ottenere il reward: condizionata all'arrivare nello stato s': $R_{s \to s'|a_j}$

$Q^{\pi}(s,a)$: secondo termine

$$E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^{k+1}r_{t+k+2}\,|s_t = s, a_t = a\right\}$$

$$P_{s\to s'|a} \triangleq \Pr(s_{t+1} = s'|s_t = s, a_t = a)$$



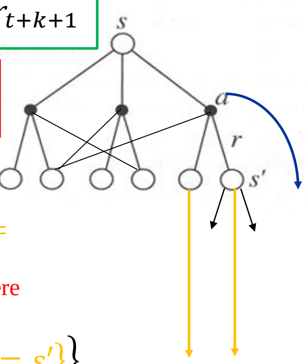
$$E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^{k+1}r_{t+k+2}\,|s_t = s, a_t = a\right\}$$

$$= \gamma\sum_{s'} P_{s\to s'|a}\, E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^{k}r_{t+k+2}\,|s_{t+1} = s'\right\}$$

Putting all together

$$\boxed{Q^{\pi}(s_t, a_t) = E_{\pi}\{R_t|s_t = s, a_t = a\} = \textstyle\sum_{k=0}^{\infty}\gamma^{k}r_{t+k+1}}$$

$$\boxed{Q^{\pi}(s_{t+1}, a_{t+1}) = E_{\pi}\{R_t|s_{t+1} = s', a_{t+1} = a'\}}$$



$$\textstyle\sum_{s'} P_{s\to s'|a} R_{s,s',a} +$$
$$= \gamma \textstyle\sum_{s'} P_{s\to s'|a}\, E_{\pi}\, \{\sum_{k=0}^{\infty}\gamma^{k}r_{t+k+2}\,|s_{t+1} = s'\} =$$

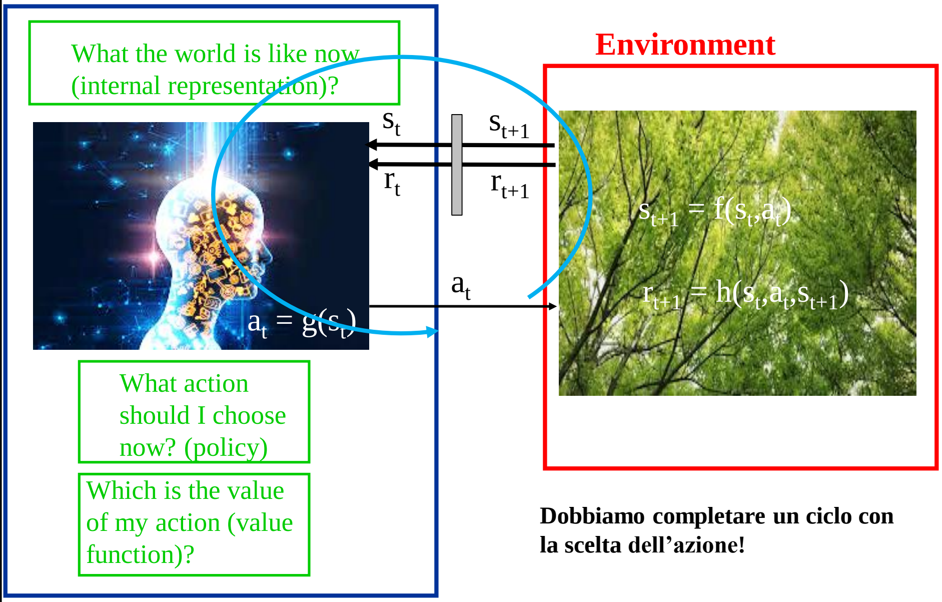
Not yet there

$$\textstyle\sum_{s'} P_{s\to s'|a}\left\{R_{s,s',a} + \gamma E_{\pi}\{\sum_{k=0}^{\infty}\gamma^{k}r_{t+k+2}\,|s_{t+1} = s'\}\right\}$$

Io termine
(a un passo)

IIo termine
(passi futuri)

Un ciclo di interazione

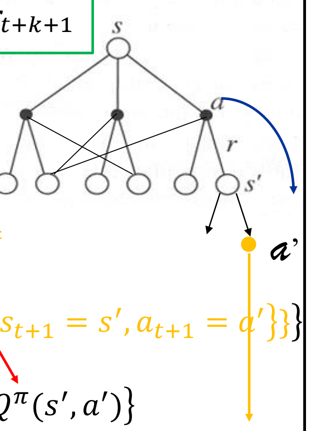


Dobbiamo completare un ciclo con la scelta dell'azione!

Formulazione ricorsiva – policy deterministica

$$Q^\pi(s_t, a_t) = E_\pi\{R_t | s_t = s, a_t = a\} = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

$$Q^\pi(s_{t+1}, a_{t+1}) = E_\pi\{R_t | s_{t+1} = s', a_{t+1} = a'\}$$



$$\sum_{s'} P_{s\to s'|a} R_{s,s',a} +$$

$$= \gamma \sum_{s'} P_{s\to s'|a}\, E_\pi \left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \,|\, s_{t+1} = s'\right\} =$$

$$\sum_{s'} P_{s\to s'|a}\left\{R_{s,s',a} + \gamma P_{a'|s'}\left\{E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \,|\, s_{t+1} = s', a_{t+1} = a'\right\}\right\}\right\}$$

$$Q^\pi(s_t, a_t) = \sum_{s'} P_{s\to s'|a}\left\{R_{s,s',a} + \gamma Q^\pi(s', a')\right\}$$

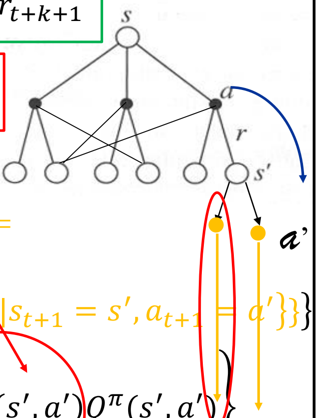
Io termine
(a un passo)

IIo termine
(passi futuri, per ogni azione a_{t+1})

Formulazione ricorsiva – policy stocastica

$$Q^\pi(s_t,a_t) = E_\pi\{R_t|s_t=s,a_t=a\} = \sum_{k=0}^{\infty}\gamma^k r_{t+k+1}$$

$$Q^\pi(s_{t+1},a_{t+1}) = E_\pi\{R_t|s_{t+1}=s',a_{t+1}=a'\}$$



$$\sum_{s'} \mathrm{P}_{s\to s'|a} R_{s,s',a} +$$

$$= \gamma\sum_{s'} \mathrm{P}_{s\to s'|a}\, E_\pi\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+2}\, |s_{t+1}=s'\right\}=$$

$$\sum_{s'} \mathrm{P}_{s\to s'|a}\left\{R_{s,s',a} + \gamma \mathrm{P}_{a'|s'}\left\{E_\pi\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+2}|s_{t+1}=s',a_{t+1}=a'\right\}\right\}\right\}$$

$$Q^\pi(s_t,a_t) = \sum_{s'} \mathrm{P}_{s\to s'|a}\left\{R_{s,s',a} + \gamma\sum_{a'}\pi(s',a')Q^\pi(s',a')\right\}$$

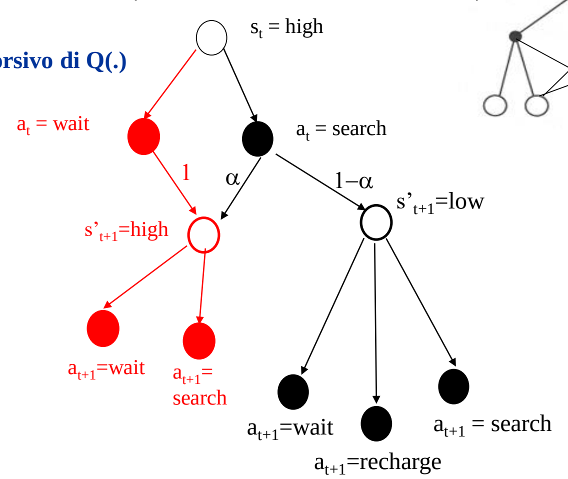
Io termine
(a un passo)

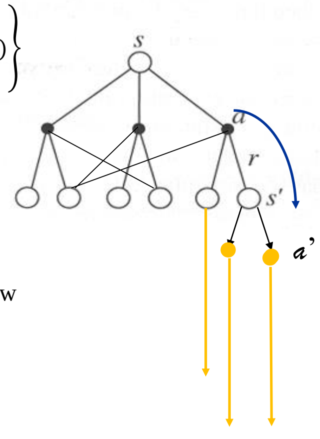
IIo termine
(passi futuri, per ogni azione a_{t+1})

Equazioni di Bellman

$$Q^\pi(s_t,a_t) = \sum_{s'} \mathrm{P}_{s\to s'|a}\left\{R_{s\to s'|a} + \gamma\sum_{a'}\pi(s',a')Q^\pi(s',a')\right\}$$

Calcolo ricorsivo di Q(.)

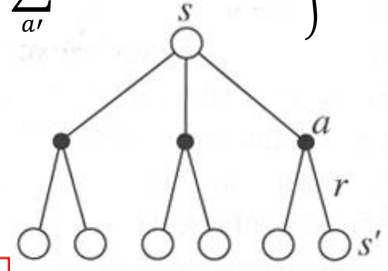




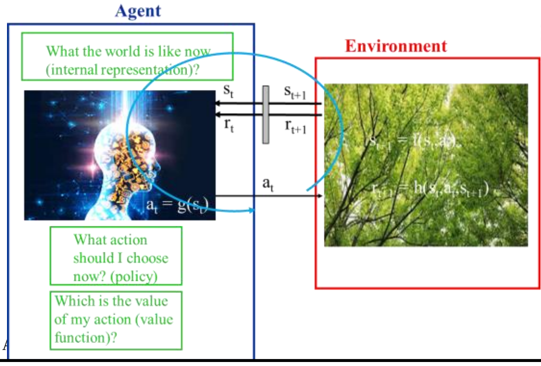
$$Q^\pi(high,wait) = R^{wait} + 0.5\,\gamma\, Q^\pi(high,wait) + 0.5\gamma Q^\pi(high,search)$$

Osservazioni

$$Q^{\pi}(s_t, a_t) = \sum_{s'} \mathrm{P}_{s \rightarrow s'|a} \left\{ R_{s,s',a} + \gamma \sum_{a'} \pi(s',a') Q^{\pi}(s',a') \right\}$$

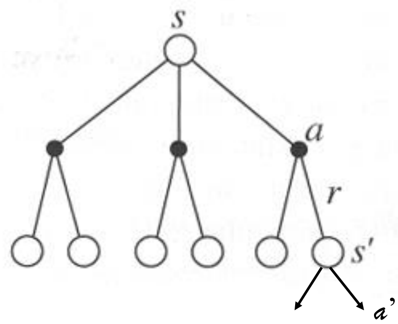


Calcolo ricorsivo di Q(.)



Passo da t a t+1 poi guardo backwards in time

Tecnica full-back



Back-up

π(s,a) fissata

t+1

Conosciamo $Q(s_t,a_t)$ $\forall s_t,a_t$ anche per $\{s'_{t+1}, a'_{t+1}\}$ quindi:

- Analizziamo la transizione da $\{s_t,a_t\}$ -> $\{s'_{t+1},a'_{t+1}\}$
- Calcoliamo un nuovo valore di Q per {s,a: $Q(s_t, a_t)$ congruente con: $Q(s_t, a_t)$ ed r_{t+1}

Full backup se esaminiamo tutti gli s' e a' (cf. DP).
Da {s', a'} mi guardo indietro e aggiorno Q(s,a).

π fissata

Meccanismo di apprendimento nel RL

Inizializzazione: se l'agente non agisce sull'ambiente non succede nulla. Occorre specificare una policy iniziale.

Ciclo dell'agente (le tre fasi sono sequenziali):

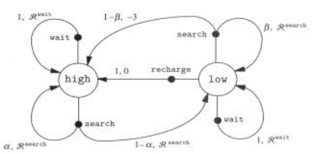
1) Implemento una policy ($\pi(s,a)$)
2) Aggiorno la Value function ($Q^{\pi}(s,a)$)
3) Aggiorno la policy.

http:\\borghese.di.unimi.it\

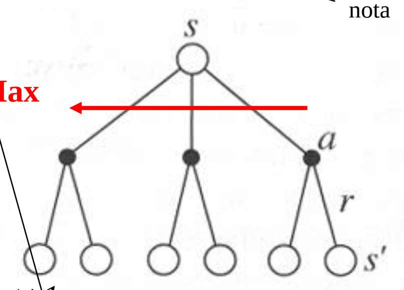
Q(s,a) – Osservazioni

$$Q^{\pi}(s_t, a_t) = \sum_{s'} P_{s \to s'|a} \left\{ R_{s,s',a} + \gamma \sum_{a'} \pi(s', a') Q^{\pi}(s', a') \right\}$$

Policy nota



Max



Per ogni stato devo valutare con informazioni esclusivamente racchiuse in 1 passo l'azione migliore a lungo termine

$t+1$

$$a_{new} : \max_a Q(s,a)$$

E' supposto noto il funzionamento dell'ambiente (simulazione)

Sommario

Le equazioni di Bellman

Differenze temporali

Q(s,a) – Osservazioni

$$Q^{\pi}(s_t, a_t) = \sum_{s'} \mathrm{P}_{s \to s' | a} \left\{ R_{s,s',a} + \gamma \sum_{a'} \pi(s', a') Q^{\pi}(s', a') \right\}$$

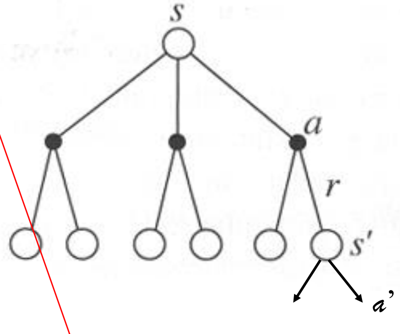
? ?

Policy nota

Per ogni stato devo valutare con informazioni esclusivamente racchiuse in 1 passo l'azione migliore a lungo termine

$$a_{new} : \max_{a} Q(s,a)$$

Cosa cambia?



Non è noto il funzionamento dell'ambiente (interazione)

Background su Temporal Difference (TD) Learning

Al tempo t abbiamo a disposizione:

$r_{t+1} = r'$ estratto (sampled) dalla distribuzione statistica: $R_{s \to s'|a_j}$

$s_{t+1} = s'$ estratto (sampled) dalla distribuzione statistica: $P_{s \to s'|a_j}$

Dopo la realizzazione di un evento, l'incertezza statistica scompare.
1 Reward certo
1 Transizione certa
vengono forniti dall'ambiente

Come si possono utilizzare per apprendere?

Confronto con il rinforzo classico

$$Q_{k+1} = Q_k - \frac{Q_k}{N_{k+1}} + \frac{r_{k+1}}{N_{k+1}} = \boxed{Q_k + \alpha[r_{k+1} - Q_k]}$$

Occupazione di memoria minima: Solo Q_k e k.
NB N_k è il numero di volte in cui è stata scelta a_j.

Questa forma è la base del RL. La sua forma generale è:

NewEstimate = OldEstimate + StepSize [Target – OldEstimate]
*NewEstimate = OldEstimate + StepSize * Error.*

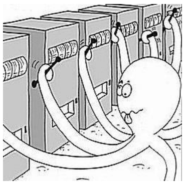
StepSize = α = 1/(N+1)	*a = cost*
Rewards weight w = 1	*Weight of i-th reward at time k: w = $(1-a)^{k-i}$*

Qual è la differenza introdotta dall'approccio che prevede comportamenti (catene di azioni)?

Un possibile aggiornamento di Q(s,a)

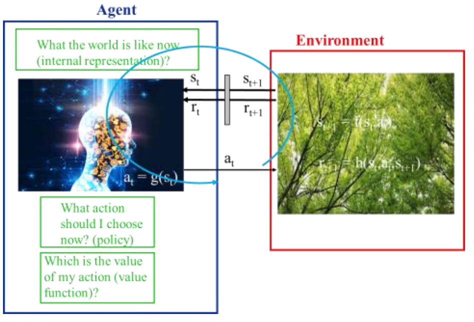
$$Q_{k+1}(a) = Q_k(a) - \frac{Q_k(a)}{N_{k+1}(a)} + \frac{r_{k+1}(a)}{N_{k+1}(a)} = Q_k(a) + \alpha[r_{k+1}(a) - Q_k(a)] =$$

$$Q_{k+1}(a) = Q_k(a) + \alpha \Delta Q_k(a)$$



Come passo ai comportamenti?

$$Q^{\pi}_{k+1}(s,a) = Q^{\pi}_k(s,a) \ + \alpha \Delta Q_k(s,a)$$

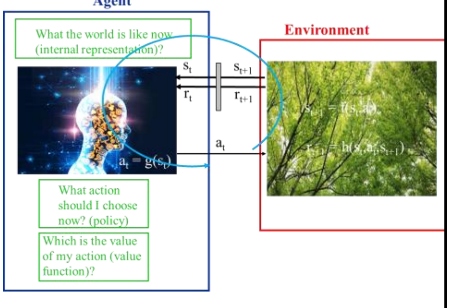


Come calcolo ΔQ_k?

Calcolo di ΔQ_k

$$Q_{k+1}(a) = Q_k(a) + \alpha[r_{k+1}(a) - Q_k(a)] =$$

$$Q_{k+1}(a) = Q_k(a) + \alpha \Delta Q_k(a)$$



Al tempo t abbiamo a disposizione:

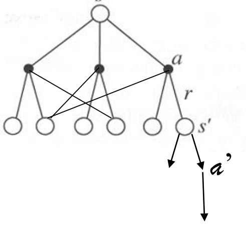
$r_{t+1} = r'$ da: $R_{s \to s'|a_j}$

$s_{t+1} = s'$ da: $P_{s \to s'|a_j}$

Quale semantica hanno Q(s,a) e r(s,a,s') nel caso dei comportamenti?

$$\boxed{Q^{\pi}_{k+1}(s,a) = Q^{\pi}_k(s,a) + \alpha\left[r' + \gamma Q^{\pi}_k(s',a') - Q^{\pi}_k(s,a)\right] =}$$

$$Q_{k+1}(a) = Q_k \ (a) + \alpha \Delta Q_k(a)$$



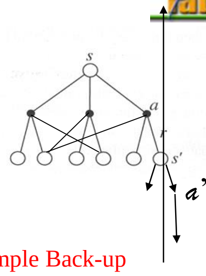
Reward a 1 passo

Reward a lungo termine da s'

TD(0) update

Ad ogni istante di tempo di ogni trial aggiorno la Value function:

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



Sample Back-up

Conosciamo $Q(s_t,a_t)$ $\forall s_t, a_t$ anche per $\{s'_{t+1}, a'_{t+1}\}$ quindi:

- Analizziamo la transizione da $\{s_t,a_t\}$ -> $\{s'_{t+1},a'_{t+1}\}$
- Calcoliamo un nuovo valore di Q per {s,a: $Q(s_t, a_t)$ congruente con: $Q(s_t, a_t)$ ed r_{t+1}

Sample backup se esaminiamo una sola coppia di s' e a' (cf. DP asincrona).
Da {s', a'} mi guardo indietro e aggiorno Q(s,a).
Percorro un solo ramo dell'albero, alla volta.

Per α che diminuisce con l'apprendimento, per k → ∞,
$Q_k^{\pi}(s,a)$ converge al valore vero di Q^{π}(s,a)

$\boxed{\pi(s,a)\ \text{fissata}}$

Posso ragionare a un passo per calcolare Q^{π}(s,a)

Confronto con il setting associativo

$$Q_{k+1} = Q_k - \frac{Q_k}{N_{k+1}} + \frac{r_{k+1}}{N_{k+1}} = \boxed{Q_k + \alpha\left[r_{k+1} - Q_k\right]}$$

Occupazione di memoria minima: Solo Q_k e k.
NB k è il numero di volte in cui è stata scelta a_j.

Questa forma è la base del RL. La sua forma generale è:

NewEstimate = OldEstimate + StepSize [Target – OldEstimate]
*NewEstimate = OldEstimate + StepSize * Error.*

StepSize = α = 1/N_{k+1}   *a = cost*

Setting α value

$\alpha(s_t, a_t, s_{t+1}) = \frac{1}{N(s_t,a_t,s_{t+1})}$, where $N(s_t, a_t, s_{t+1})$ represents the number of occurrences of s_t, a_t, s_{t+1}. With this setting the estimated Q tends to the expected value of Q(s,a).

Per semplicità si assume solitamente $\alpha < 1$ costante. In questo caso, Q(s,a) assume il valore di una media pesata dei reward a lungo termine collezionati a partire da (s,a), con peso: $(1-\alpha)^k$: *exponential recency-weighted average*.

α che decresce dolcemente a zero consente la convergenza del Sistema stocastico.

Esempio

Stima del tempo di percorrenza da casa all'ufficio su un percorso ben definite (policy deterministica).

La durata dei diversi segmenti può variare da giorno a giorno e quindi la stima della durata totale viene corretta conseguentemente.

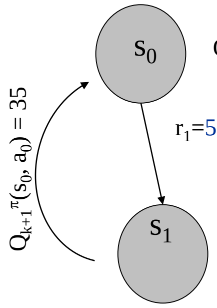
La stima corrente del tempo totale è data dalla somma dei tempi per:

- Dall'ufficio al parcheggio: 5 minuti (time to go = 35 minuti)
- Dal parcheggio all'uscita dell'autostrada: 15 minuti (time to go = 30 minuti)
- Dall'uscita dell'autostrada alla strada di casa: 5 minuti (time to go = 15 minuti)
- Dalla strada di casa a casa: 7 minuti (time to go = 10 minuti)
- Dal parcheggio a casa: 3 minuti (time to go = 3 minuti)

- In totale 35 minuti.

Learning $Q^\pi(s,a)$ - I

s_0 = ufficio; Q_k^π(s_0, vado_parcheggio) = 35 *minuti*; Q_{k+1}^π(s_0, vado_parcheggio) = *35 minuti* (potrei fare altre scelte, e.g. andare alla metropolitana, ma la policy prescrive di andare a prendere l'auto nel parcheggio perchè era considerata la soluzione più veloce).



Q_k^π(s_0,vado_parcheggio) = 35

r_1=5 minuti per raggiungere il parcheggio

Q_k^π(s_1, imbocco_autostrada) = 30

Raggiungo il parcheggio (stato s_1). Vedo che piove e spero che il tempo totale non vari di molto. Stimo il tempo per arrivare a casa in 30 minuti: all'uscita del parcheggio imbocco l'autostrada Q_k^π(s_1. imbocco l'autostrada)

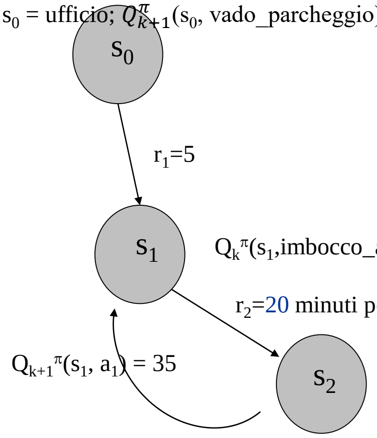
Aggiorno il tempo totale, ovverosia il tempo dallo stato s_0:

$$Q_{k+1}^\pi(s_0,a) = Q_k^\pi(s_0,a) + \alpha\left[r' + \gamma Q_k^\pi(s_1,a') - Q_k^\pi(s_0,a)\right] = 35+[5+30-35] = 35$$

Suppongo $\alpha = \gamma = 1$

Learning $Q^\pi(s,a)$ - II

s_1 = parcheggio; Q_k^π(s_1, imbocco_autostrada) = 30 *minuti*;
Q_{k+1}^π(s_1, imbocco_autostrada = 30 *minuti*;
(potrei fare altre scelte, e.g. tornare in ufficio; una volta scelto di uscire, aggiorno il valore dell'azione uscire dal parcheggio, quando sono nel parcheggio)
s_0 = ufficio; Q_{k+1}^π(s_0, vado_parcheggio) = 40 *minuti*



Esco dall'autostrada (stato s_2). Sull'autostrada c'era traffico più lento del solito, impiego 20 minuti, 5 minuti in più del solito.

Q_k^π(s_1,imbocco_autostrada) = 30

r_2=20 minuti per percorrere l'autostrada

Q_k^π(s_2,esco_autostrada) = 15

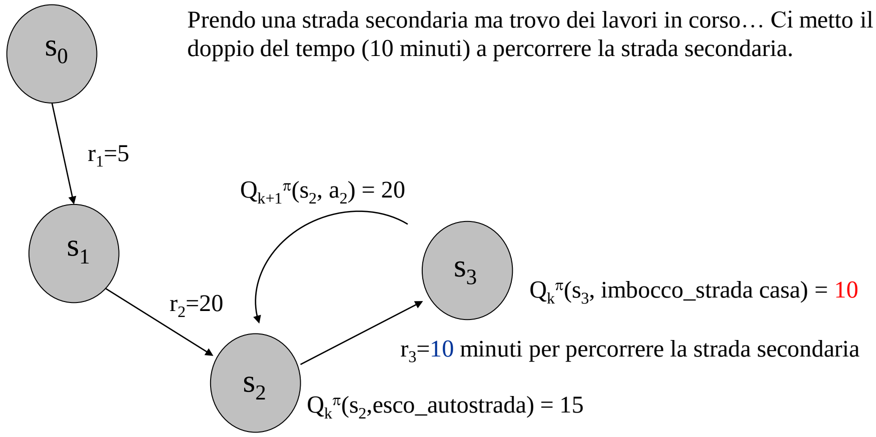
Aggiorno il tempo totale dallo stato s_1:

$$Q_{k+1}^\pi(s_1,a) = Q_k^\pi(s_1,a) + \alpha\left[r' + \gamma Q_k^\pi(s_2,a') - Q_k^\pi(s_1,a)\right] = 30+[20+15-30] = 35$$

Learning $Q^{\pi}(s,a)$ - III

s_2 = esco_autostrada; $Q_k^{\pi}(s_2$, esco_autostrada = 15 min; Q_{k+1}^{π} (s_2, esco_autostrada) = 20 min;
s_0 = ufficio; $Q_{k+1}^{\pi}(s_0$, vado_parcheggio) = 45 $minuti$

Prendo una strada secondaria ma trovo dei lavori in corso… Ci metto il doppio del tempo (10 minuti) a percorrere la strada secondaria.



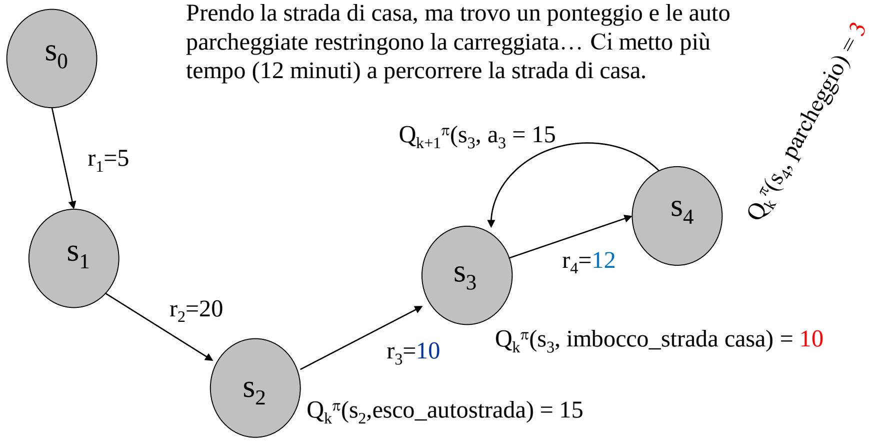
Aggiorno il tempo totale dallo stato s_2:

$Q_{k+1}^{\pi}(s_2,a) = Q_k^{\pi}(s_2,a) + \alpha\left[r' + \gamma Q_k^{\pi}(s_3,a') - Q_k^{\pi}(s_2,a)\right]$ = 15+[10+10-15] = 20

Learning $Q^{\pi}(s,a)$ - IV

s_3 = imbocco_strada casa; $Q_k^{\pi}(s_3$, imbocco_strada casa) = 10 $min;$
$Q_{k+1}^{\pi}(s_3$, imbocco_strada casa) = 15 min;
s_0 = ufficio; $Q_{k+1}^{\pi}(s_0$, vado_parcheggio) = 43 $minuti$

Prendo la strada di casa, ma trovo un ponteggio e le auto parcheggiate restringono la carreggiata… Ci metto più tempo (12 minuti) a percorrere la strada di casa.

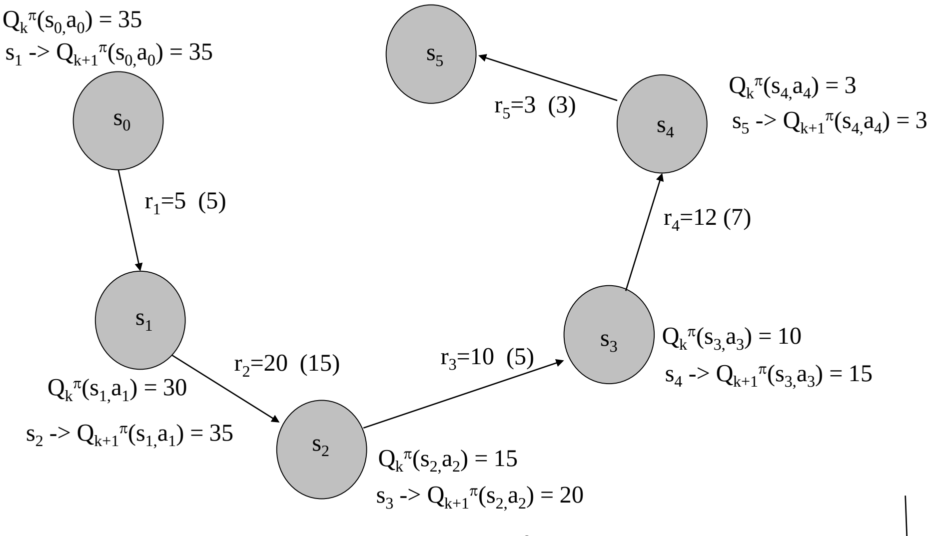


Aggiorno il tempo totale dallo stato s_2:

$Q_{k+1}^{\pi}(s_2,a) = Q_k^{\pi}(s_2,a) + \alpha\left[r' + \gamma Q_k^{\pi}(s_3,a') - Q_k^{\pi}(s_2,a)\right]$ = 10+[12+3-10] = 15

Learning $Q^{\pi}(s,a)$

s_0 = ufficio; s_5 = casa.



Come i diversi reward istantanei modificano Q^{π} (s,a)?

Esempio – dopo il primo trial

Stima del tempo di percorrenza da casa all'ufficio su un percorso ben definite (policy deterministica).

La durata dei diversi segmenti può variare da giorno a giorno e quindi la stima della durata totale è stata corretta conseguentemente all'esplorazione.

La stima corrente del tempo totale è data dalla somma dei tempi per:

- Dall'ufficio al parcheggio: 5 minuti (time to go = 35 minuti)
- Dal parcheggio all'uscita dell'autostrada: 15 minuti (time to go = 35 minuti)
- Dall'uscita dell'autostrada alla strada di casa: 5 minuti (time to go = 20 minuti)
- Dalla strada di casa a casa: 7 minuti (time to go = 15 minuti)
- Dal parcheggio a casa: 3 minuti (time to go = 3 minuti)

Si sono create diverse incongruenze (ad esempio il time to go è di 35 minuti dall'ufficio come dal parcheggio!), che verranno corrette via via che si ripeteranno le stesse situazioni.

Attualmente la stima agiornata di Q(.) è per lo stato prima di quello finale ed è di 3 minuti. La stima di Q(.) per gli stati precedenti, viene via via aggiornata nei trial successivi.

Ruolo di α

$Q_{k+1}(s_1,a_1) = Q_k(s_1,a_1) + \alpha\,(r_1 + \gamma Q(s_2,a_2) - Q(s_2,a_2)) = 30+\alpha\,(20+15-30) = 30+\alpha*5$

Stima iniziale del tempo di percorrenza dal parcheggio: 30m
Tempo per percorrere l'autostrada: 20m
Stima del tempo di percorrenza dall'uscita del parcheggio: 35min (per $\alpha = 1$)

$\alpha < 1$.

If $\alpha \ll 1$ aggiorno molto lentamente la value function.

If $\alpha = 1/k(s,a)$ aggiorno la value function in modo da tendere al valore atteso. Devo memorizzare le occorrenze della coppia stato-azione s,a.

If α = cost. Aggiorno la value function, pesando maggiormente i risultati collezionati dalle visite dello stato più recenti.

La convergenza è garantita per α che decresce gradualmente verso zero.

Sommario

Le equazioni di Bellman

Differenze temporali