

# Sistemi Intelligenti Reinforcement Learning: Equazioni di Bellman

Alberto Borghese

Università degli Studi di Milano  
Laboratorio di Sistemi Intelligenti Applicati (AIS-Lab)  
Dipartimento di Scienze dell'Informazione  
[borgnese@dsi.unimi.it](mailto:borgnese@dsi.unimi.it)



A.A. 2006-2007

1/41

<http://homes.dsi.unimi.it/~borgnese/>



## Sommario



Determinazione della value function. Esempio.

Le equazioni di Bellman

Value function ottima

A.A. 2006-2007

2/41

<http://homes.dsi.unimi.it/~borgnese/>



# Reinforcement Learning Problem



Given: Repeatedly...

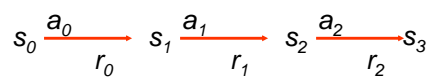
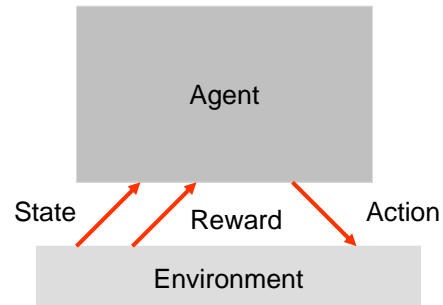
- Executed action
- Observed state
- Observed reward

Learn action policy  $\pi: S \rightarrow A$

- ◆ Maximizes life reward  
 $r_0 + \gamma r_1 + \gamma^2 r_2 \dots$   
from any start state.
- ◆ Discount:  $0 < \gamma < 1$

Note:

- Unsupervised learning
- Delayed reward



Goal: **Learn** to choose actions that maximize life reward

$$r_0 + \gamma r_1 + \gamma^2 r_2 \dots$$



# 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.

### Policy:

$A(s = \text{high}) = \{\text{Search, Wait}\}$

$A(s = \text{low}) = \{\text{Search, Wait, Recharge}\}$

**Goal:** collezionare il maggior numero di lattine.



## Funzionamento del Robot



### Funzione Stato prossimo:

$$P_{s \rightarrow s'|a} = \Pr\{s_{t+1} = s' | s_t = s, a_t = a\}$$

Se il livello di energia è alto ( $s_t = \text{alto}$ ):

se scelgo Wait -  $s_{t+1} = \text{alto}$ .

se scelgo Search,  $s_{t+1}$  avrà una certa probabilità di diventare low.

$$P_{\text{high} \rightarrow \text{low} | \text{Search}} = \Pr\{s_{t+1} = \text{low} | s_t = \text{high}, a_t = \text{Search}\} = \alpha$$

Se il livello di energia è basso ( $s_t = \text{basso}$ ):

se scelgo Wait -  $s_{t+1} = \text{basso}$ .

se scelgo Recharge -  $s_{t+1} = \text{alto}$ .

se scelgo Search,  $s_{t+1}$  avrà una certa probabilità di fermarsi.

$$P_{\text{low} \rightarrow \text{low} | \text{Search}} = \Pr\{s_{t+1} = \text{low} | s_t = \text{low}, a_t = \text{Search}\} = \beta$$

A.A. 2006-2007

5/41

<http://homes.dsi.unimi.it/~borghese/>



## Reward del Robot



### Funzione Reward:

$$R_{s \rightarrow s'|a} = E\{r_{t+1} = r' | s_t = s, a_t = a, s_{t+1} = s'\}$$

$R^{\text{search}}$  reward se il robot sta cercando.

$R^{\text{wait}}$  reward se il robot sta cercando.

-3 se occorre portarlo a ricaricarsi.

0 se il robot va autonomamente a ricaricarsi.

$$R^{\text{search}} > R^{\text{wait}}$$

A.A. 2006-2007

6/41

<http://homes.dsi.unimi.it/~borghese/>



## Forma tabellare

s	s'	a	$P_{s \rightarrow s'} a$	$R_{s \rightarrow s'} a$
alta	alta	ricerca	$\alpha$	$R^{\text{search}}$
alta	bassa	ricerca	$1 - \alpha$	$R^{\text{search}}$
bassa	alta	ricerca	$1 - \beta$	-3
bassa	bassa	ricerca	$\beta$	$R^{\text{search}}$
alta	alta	attesa	1	$R^{\text{wait}}$
alta	bassa	attesa	0	$R^{\text{wait}}$
bassa	alta	attesa	0	$R^{\text{wait}}$
bassa	bassa	attesa	1	$R^{\text{wait}}$
bassa	alta	ricarica	1	0
bassa	bassa	ricarica	0	0
alta	Non esiste	ricarica	X	X

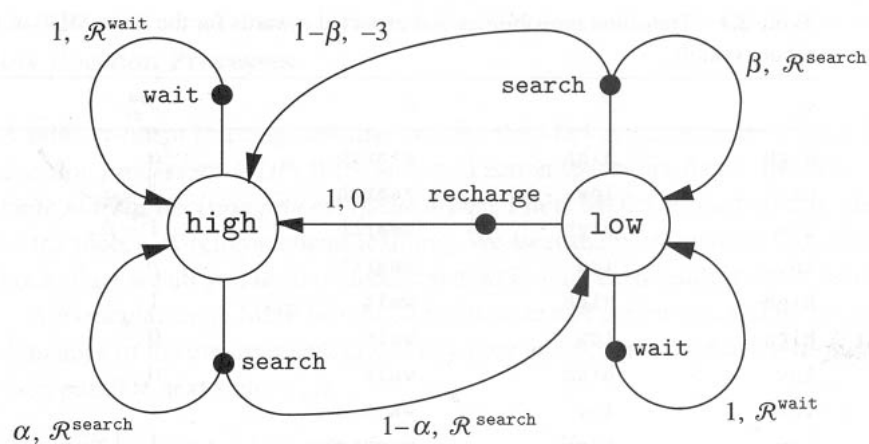
A.A. 2006-2007

7/41

<http://homes.dsi.unimi.it/~borghese/>



## State Transition Graph



Probabilistic Finite State Machine

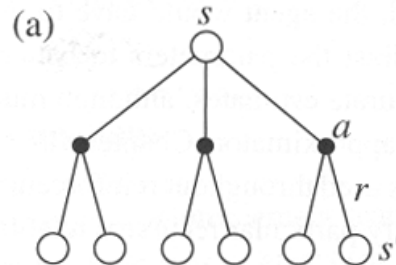
A.A. 2006-2007

8/41

<http://homes.dsi.unimi.it/~borghese/>



## Policy



La policy deve essere ancora determinata. Come fa l'agente a determinare la policy ottimale?

Archi multipli fuoriuscenti da un'azione sono associati alla probabilità di scegliere quel cammino (ambiente stocastico).

Archi multipli fuoriuscenti da uno stato, sono associati alla policy.

A.A. 2006-2007

9/41

<http://homes.dsi.unimi.it/~borghese/>



## Value function & policy

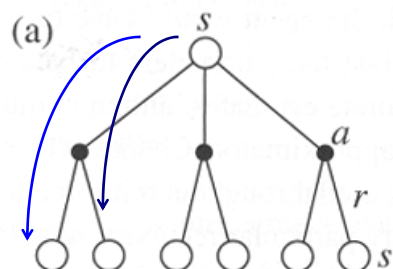
Nulla è detto sulla policy: dato uno stato, in quale nodo azione mi sposto?

Vogliamo costruire agenti lungimiranti

State-Value function  
(function of the policy):

$$V^\pi(s) = E_\pi \{ R_t \mid s_t = s \} =$$

$$E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\}$$



Massimizzo la ricompensa a lungo termine,  $V(\cdot)$ . Dipende dalla policy:

A.A. 2006-2007

10/41

<http://homes.dsi.unimi.it/~borghese/>



# Value function e modelli markoviani

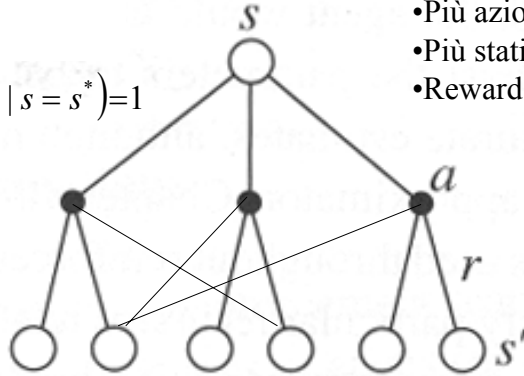


Anche la policy può essere stocastica.

Per ogni stato devo valutare:

- Più azioni.
- Più stati prossimi
- Reward stocastici.

$$\sum_{j=1}^{N_{\text{azioni}}} \Pr(a_j | s = s^*) = 1$$



$$\sum_{k=1}^{N_{\text{stati}}} \Pr(s_{t+1} = s_k | s_t = s'; a_t = a_j) = 1$$

$$V^\pi(s) = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\}$$

A.A. 2006-2007

11/41

<http://homes.dsi.unimi.it/~borghese/>



# Esempio di calcolo della Value function



Value function

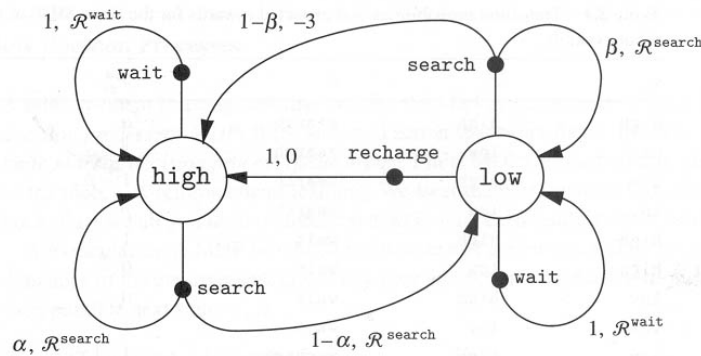
V(high) = ?

V(low) = ?

Policy

a(high) = ?

a(low) = ?



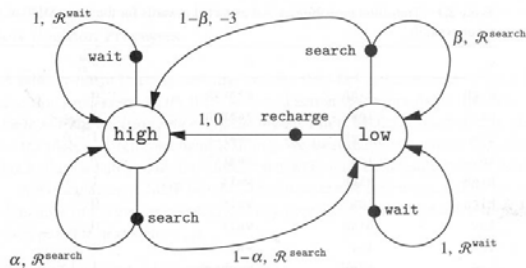
A.A. 2006-2007

12/41

<http://homes.dsi.unimi.it/~borghese/>



## Esempio di calcolo della funzione valore



$\alpha=0.4, \beta=0.1, \gamma=0.8, R_{search}=3, R_{wait}=1$

$$V_h = \Pr(W) \times 1 \times [1 + 0.8V_h] + \Pr(S) \times 0.4 \times [3 + 0.8V_h] + \Pr(S) \times 0.6 \times [3 + 0.8V_l]$$

*Wait*
*Search -> high*
*Search -> low*

$$V_l = \Pr(W) \times 1 \times [1 + 0.8V_l] + \Pr(S) \times 0.1 \times [3 + 0.8V_l] + \Pr(S) \times 0.9 \times [-3 + 0.8V_h] + \Pr(R) \times 1 \times [0 + 0.8V_h]$$

*Wait*
*Recharge*
*Search -> High*

*Search -> Low*

**Sistema lineare di 2 equazioni nelle 2 incognite:  $V_h$  e  $V_l$**   
**Come calcolo  $V_h$  e  $V_l$ ?**

A.A. 2006-2007

13/41

<http://homes.dsi.unimi.it/~borghese/>



## Esempio di calcolo della funzione valore - I



$$V_h = \Pr(W) \times 1 \times [1 + 0.8V_h] + \Pr(S) \times 0.4 \times [3 + 0.8V_h] + \Pr(S) \times 0.6 \times [3 + 0.8V_l]$$

*Wait*
*Search -> high*
*Search -> low*

$$V_l = \Pr(W) \times 1 \times [1 + 0.8V_l] + \Pr(S) \times 0.1 \times [3 + 0.8V_l] + \Pr(S) \times 0.9 \times [-3 + 0.8V_h] + \Pr(R) \times 1 \times [0 + 0.8V_h]$$

*Wait*
*Recharge*
*Search -> High*

*Search -> Low*

Devo specificare una policy (stocastica):

$$s = \text{high} \quad [\Pr(W) = 0.4 \quad \Pr(S) = 0.6]$$

$$s = \text{low} \quad [\Pr(W) = 0.4 \quad \Pr(S) = 0.5 \quad \Pr(R) = 0.1]$$

$$V_h = 0.4 \times 1 \times [1 + 0.8V_h] + 0.6 \times 0.4 \times [3 + 0.8V_h] + 0.6 \times 0.6 \times [3 + 0.8V_l]$$

$$V_l = 0.4 \times 1 \times [1 + 0.8V_l] + 0.5 \times 0.1 \times [3 + 0.8V_l] + 0.5 \times 0.9 \times [-3 + 0.8V_h] + 0.1 \times 1 \times [0 + 0.8V_h]$$

A.A. 2006-2007

14/41

<http://homes.dsi.unimi.it/~borghese/>



## Esempio di calcolo della funzione valore - II



Devo specificare una policy (stocastica):

$$s = \text{high} \quad [\text{Pr}(W) = 0.4 \quad \text{Pr}(S) = 0.6]$$

$$s = \text{low} \quad [\text{Pr}(W) = 0.4 \quad \text{Pr}(S) = 0.5 \quad \text{Pr}(R) = 0.1]$$

$$V_h = 0.4 \times 1 \times [1 + 0.8V_h] + 0.6 \times 0.4 \times [3 + 0.8V_h] + 0.6 \times 0.6 \times [3 + 0.8V_l]$$

$$V_l = 0.4 \times 1 \times [1 + 0.8V_l] + 0.5 \times 0.1 \times [3 + 0.8V_l] + 0.5 \times 0.9 \times [-3 + 0.8V_h] + 0.1 \times 1 \times [0 + 0.8V_h]$$

$$0.488 V_h = 2.20 + 0.288 V_l \quad \Rightarrow \quad V_h \approx 6,35$$

$$0.64V_l = -0.8 + 0.44 V_h \quad \Rightarrow \quad V_l \approx 3,12$$

Cosa si può concludere? Quasi nulla.



## Esempio di calcolo della funzione valore - III



Devo specificare una policy (stocastica), robot più attivo:

$$s = \text{high} \quad [\text{Pr}(W) = 0.2 \quad \text{Pr}(S) = 0.8]$$

$$s = \text{low} \quad [\text{Pr}(W) = 0.2 \quad \text{Pr}(S) = 0.7 \quad \text{Pr}(R) = 0.1]$$

$$V_h = 0.2 \times 1 \times [1 + 0.8V_h] + 0.8 \times 0.4 \times [3 + 0.8V_h] + 0.8 \times 0.6 \times [3 + 0.8V_l]$$

$$V_l = 0.2 \times 1 \times [1 + 0.8V_l] + 0.7 \times 0.1 \times [3 + 0.8V_l] + 0.7 \times 0.9 \times [-3 + 0.8V_h] + 0.1 \times 1 \times [0 + 0.8V_h]$$

$$0.488 V_h = 2.20 + 0.288 V_l \quad \Rightarrow \quad V_h \approx -14.2$$

$$0.64V_l = -0.8 + 0.44 V_h \quad \Rightarrow \quad V_l \approx -28.465$$

E' una policy peggiore.

**Come utilizziamo V per determinare la policy ottimale?**





## Sommario



Determinazione della value function. Esempio.

**Le equazioni di Bellman**

Value function ottima



## Il modello markoviano



Il comportamento dell'ambiente è definito dallo stato:  $S = \{s_j\}$

Per ogni stato l'agente sceglie un'azione:  $a = a(s)$      $A = \{a_k\}$

**Policy di un agente:  $\pi(s, a)$  è quanto dobbiamo definire.**

L'ambiente ha una evoluzione stocastica rappresentata da un MDP:

$$P_{s_t=s \rightarrow s_{t+1}=s' | a_t=a} = \Pr\{s_{t+1} = s' | s_t = s, a_t = a\}$$

Inoltre, ad ogni istante fornisce un reward immediato associato alla transizione, stimato all'istante t come:

$$R_{s_t=s \rightarrow s_{t+1}=s' | a_t=a} = E\{r_{t+1} = r' | s_t = s, a_t = a, s_{t+1} = s'\}$$

$$\forall s \in S; \forall a \in A$$



## La value function

Nulla è detto sulla policy: dato uno stato, quale azione scegliere? In quale stato mi sposto?

Vogliamo costruire agenti lungimiranti.

### State-Value function:

$$V^\pi(s) = E_\pi \{R_t \mid s_t = s\} = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\}$$

Massimizzo la ricompensa a lungo termine,  $V(\cdot)$ . Dipende dalla policy:

A.A. 2006-2007

19/41

<http://homes.dsi.unimi.it/~borghese/>



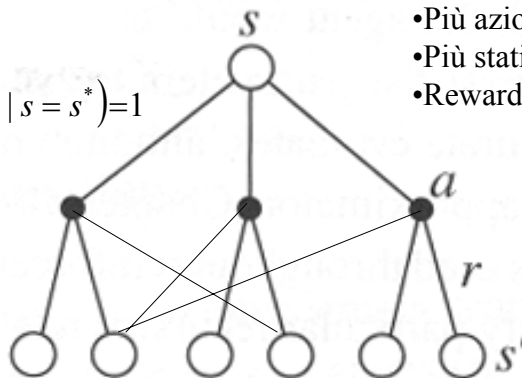
## Value function e modelli markoviani

Anche la policy può essere stocastica.

Per ogni stato devo valutare:

- Più azioni.
- Più stati prossimi
- Reward stocastici.

$$\sum_{j=1}^{N_{\text{azioni}}} \Pr(a_j \mid s = s^*) = 1$$



$$\sum_{k=1}^{N_{\text{stati}}} \Pr(s_{t+1} = s_k \mid s_t = s'; a_t = a_j) = 1$$

$$V^\pi(s) = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\}$$

A.A. 2006-2007

20/41

<http://homes.dsi.unimi.it/~borghese/>



## Calcolo ricorsivo della Value function



$$V^\pi(s) = E_\pi \{R_t | s_t = s\} = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\}$$

$$V^\pi(s') = E_\pi \{R_{t+1} | s_{t+1} = s'\}$$

Relazione?

$$V^\pi(s) = E_\pi \left\{ r_{t+1} + \sum_{k=1}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\} =$$

$$V^\pi(s) = E_\pi \left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s \right\}$$



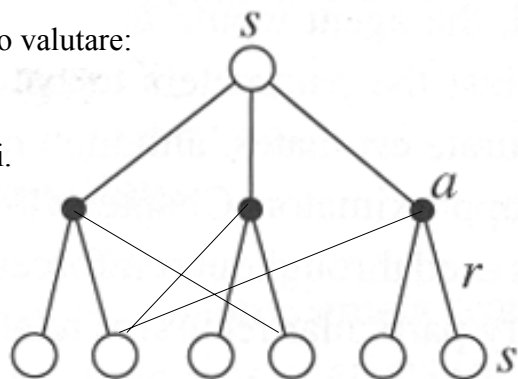
## $V^\pi(s)$ : primo termine



$$E_\pi \{r_{t+1} | s_t = s\} = \left[ \sum_{a_j} \pi(a_j, s) \right] \sum_{s'} P_{s \rightarrow s' | a_j} [R_{s \rightarrow s' | a_j}]$$

Per ogni stato devo valutare:

- Più azioni.
- Più stati prossimi
- Reward stocastici.



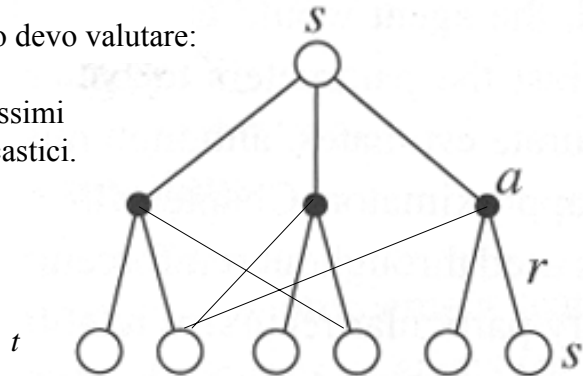


## $V^\pi(s)$ : secondo termine

$$E_\pi \left\{ \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s \right\} = \gamma \sum_t E_\pi \{ R_{t+1} \mid s_{t+1} = s'_t \} \Pr(s_{t+1} = s_k \mid s_t = s)$$

Per ogni stato devo valutare:

- Più azioni.
- Più stati prossimi
- Reward stocastici.



A.A. 2006-2007

23/41

<http://homes.dsi.unimi.it/~borghese/>



## Calcolo ricorsivo della Value function

$$V^\pi(s) = E_\pi \{ R_t \mid s_t = s \} = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\}$$

$$V^\pi(s') = E_\pi \{ R_{t+1} \mid s_{t+1} = s' \}$$

Legame?

Policy Next-state

$$P_{s \rightarrow s' | a} = \Pr \{ s_{t+1} = s' \mid s_t = s, a_t = a \}$$

$$V^\pi(s) = \sum_{a_j} \pi(a_j, s) \sum_{s'} P_{s \rightarrow s' | a_j} R_{s \rightarrow s' | a_j} + E_\pi \left\{ \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s \right\}$$

$$V^\pi(s) = \left\{ \sum_{a_j} \pi(a_j, s) \sum_{s'_l} \left[ P_{s \rightarrow s'_l | a_j} \left[ R_{s \rightarrow s'_l | a_j} + \gamma V^\pi(s'_l) \right] \right] \right\}$$

**Bellman's equation**

A.A. 2006-2007

24/41

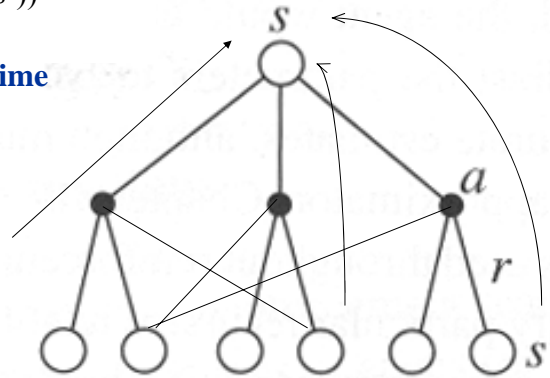
<http://homes.dsi.unimi.it/~borghese/>



## Osservazioni

$$V^\pi(s) = \text{funz}(V^\pi(s'))$$

**Backwards in time**



$$V^\pi(s) = \left\{ \sum_{a_j} \pi(a_j, s) \sum_{s_l'} \left\{ P_{s \rightarrow s_l' | a_j} \left[ R_{s \rightarrow s_l' | a_j} + \gamma V^\pi(s_l') \right] \right\} \right\}$$

A.A. 2006-2007

25/41

<http://homes.dsi.unimi.it/~borghese/>



## Sommario

Determinazione della value function. Esempio.

Le equazioni di Bellman

**Value function ottima**

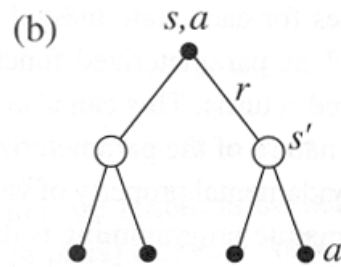
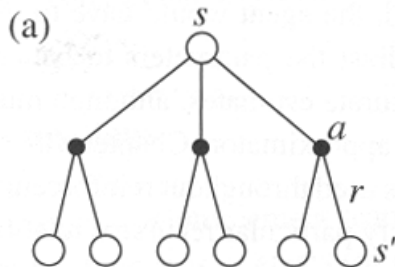
A.A. 2006-2007

26/41

<http://homes.dsi.unimi.it/~borghese/>



## Policy



La policy deve essere ancora determinata. Come fa l'agente a determinare la policy ottimale?

Archi multipli fuoriuscenti da un'azione sono associati alla probabilità di scegliere quel cammino (ambiente stocastico).

Archi multipli fuoriuscenti da uno stato, sono associati alla policy.

A.A. 2006-2007

27/41

<http://homes.dsi.unimi.it/~borghese/>



## Ottimizzazione policy



Per ogni stato scelgo le azioni secondo la policy:  $\pi(s,a)$ .

Posso ordinare la Value function  $V(s)$  in funzione delle azioni scelte in  $s$  (policy).

Si definisce una policy,  $\pi_1$ , migliore di un'altra,  $\pi_2$ , se e solo se:

$$V^{\pi_1}(s) \geq V^{\pi_2}(s) \quad \forall s.$$

In particolare si definisce una politica ottima,  $\pi^*$ , se e solo se:

$$V^*(s) \geq V^{\pi}(s) \quad \forall s$$

$$Q^*(s,a) \geq Q^{\pi}(s,a) \quad \forall [s,a]$$

A.A. 2006-2007

28/41

<http://homes.dsi.unimi.it/~borghese/>



## Riassunto

Posso determinare la Value function in modo ricorsivo. Per ogni stato, sarà funzione dell'output dell'ambiente in quell'istante (attraverso la funzione stato prossimo ed il reward istantaneo) e della policy scelta in quell'istante e dei reward a lungo termine attesi negli stati in cui l'ambiente mi porta.

Per scegliere la policy devo esaminare il reward a lungo termine che mi si prospetta nello stato in cui mi trovo e scegliere l'azione che lo massimizza.

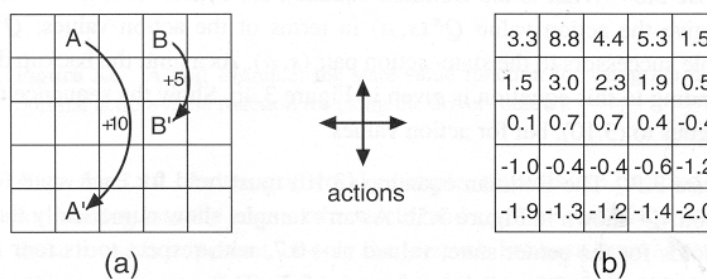
A.A. 2006-2007

29/41

<http://homes.dsi.unimi.it/~borghese/>



## Esempio (possibile progetto)



**Figure 3.5** Grid example: (a) exceptional reward dynamics; (b) state-value function for the equiprobable random policy.

$a = \pi(s)$  equiprobabile =  $0.25 \forall a$

$R(s,a)$  per stati sui bordi e azioni che portano fuori dal bordo.

$R(s)$  per tutti gli altri stati (= 0, +5, +10)

A.A. 2006-2007

30/41

<http://homes.dsi.unimi.it/~borghese/>



## Esempio di valore di uno stato



t:  $s = s_t = C$

t + 1:  $s = s' = s_{t+1}$

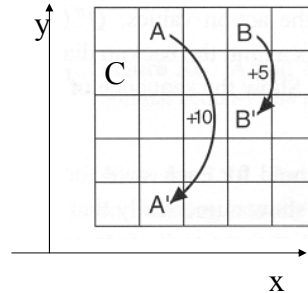
C + Dy p = 0.25 r = 0

C - Dy p = 0.25 r = 0

C + Dx p = 0.25 r = 0

C p = 0.25 r = -1

$r_{t+1} = -0.25$



t + 2, a = +Dy

→ C + Dy p = 0.25 r = -1

→ C p = 0.25 r = 0

→ A p = 0.25 r = 0

→ C + Dy p = 0.25 r = -1

$r_{t+1} = -0.50$

a = -Dy

→ C r = 0

→ C - 2Dy r = 0

→ C - Dy + Dx r = 0

→ C - Dy r = -1

$r_{t+1} = -0.25$

A.A. 2006-2007

31/41

<http://homes.dsi.unimi.it/~borghese/>



## Calcolo della funzione valore di uno stato



t:  $s = s_t = C$

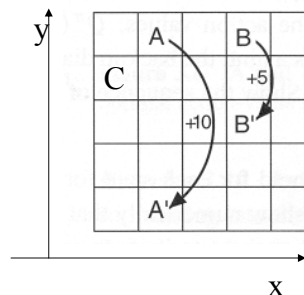
t + 1:  $s = s' = s_{t+1}$

I C + Dy p = 0.25 r = 0

II C - Dy p = 0.25 r = 0

III C + Dx p = 0.25 r = 0

IV C p = 0.25 r = -1



Per calcolare il valore di s:  $V(s)$ , devo analizzare il valore di ogni  $s'$ :

$\pi(s,a) \text{ -- } P_{s \rightarrow s' | a} \text{ -- } R_{s \rightarrow s' | a} \text{ -- } V^\pi(s') \leftarrow \text{Value of } s'$

policy

Risposta ambiente

Reward

A.A. 2006-2007

32/41

<http://homes.dsi.unimi.it/~borghese/>





## La Value function sulle azioni



La value function può riguardare le azioni.

Action-Value function

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

Massimizzo la ricompensa a lungo termine,  $Q(\cdot)$ . Dipende dalla policy.

Dove abbiamo già trovato una Value function associata alle azioni?



## Confronto con il setting non associativo



	Setting non associativo	Setting associativo
Task	Azioni	Comportamenti (catena di azioni)
Reward	Reward istantaneo	Somma (scontata) dei reward collezionati lungo il task.
Max	Reward atteso sulla singola azione	Reward del comportamento
Orizzonte temporale del task	Finito (1 azione)	Finito / infinito per il singolo task
Policy	Stocastica	Stocastica
Stato	Non definito	Markoviano



## Sommario



Determinazione della value function. Esempio.

Le equazioni di Bellman

Value function ottima