

Sistemi Intelligenti Reinforcement Learning: Q-learning

Alberto Borghese

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



A.A. 2013-2014

1/18

<http://\borghese.di.unimi.it/>



Sommario



$Q(\lambda)$

A.A. 2013-2014

2/18

<http://\borghese.di.unimi.it/>



Proprietà del rinforzo

L'ambiente o l'interazione può essere complessa.

Il rinforzo può avvenire solo dopo una più o meno lunga sequenza di azioni (**delayed reward**).

E.g. agente = giocatore di scacchi.
 ambiente = avversario.

Problemi collegati:

temporal credit assignment.
structural credit assignment.

L'apprendimento non è più da esempi, ma dall'osservazione del proprio comportamento nell'ambiente.



Formulazione di TD(0)

Correggo la stima corrente valutando l'"errore" ad un passo.

$$Q^{\pi}_{k+1}(s_t, a_t) = Q^{\pi}_k(s_t, a_t) + \alpha [r_{t+1} + \gamma Q^{\pi}_k(s_{t+1}, a_{t+1}) - Q^{\pi}_k(s_t, a_t)]$$

$$\Delta Q^{\pi}(s_t, a_t) = + \alpha \delta_k$$
$$\delta_k = [r_{t+1} + \gamma Q^{\pi}_k(s_{t+1}, a_{t+1}) - Q^{\pi}_k(s_t, a_t)]$$

Estensione dell'orizzonte temporale dell'apprendimento



Cosa rappresenta la Eligibility trace



Buffer di memoria: contiene traccia di eventi passati (stati visitati, azioni...); la traccia evapora nel tempo.

Quando viene calcolato un errore usando metodi basati su TD, la eligibility trace suggerisce quali variabili aggiornare (credit assignment).

Amplia l'orizzonte temporale sul quale fare l'aggiornamento a più di 1 passo.

Definisce se uno stato è eleggibile e "quanto" sia eleggibile.

NB L'errore modifica il valore di $Q(s,a)$ ma questo modifica a sua volta il valore di $Q(s_{t-1}, a_{t-1})$

A.A. 2013-2014

5/18

<http://borghese.di.unimi.it/>



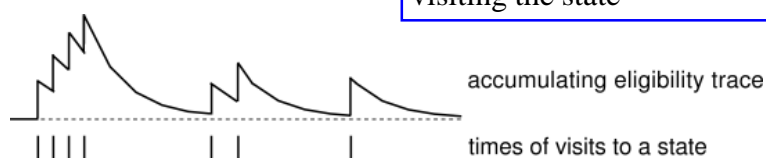
Eligibility trace per la funzione Q



$$e_t(s, a) = \begin{cases} \gamma \lambda e_{t-1}(s, a) + 1 & \text{if } s = s_t \text{ and } a = a_t; \\ \gamma \lambda e_{t-1}(s, a) & \text{otherwise.} \end{cases} \quad \text{for all } s, a$$

decay

Increases: depends only on visiting the state



$$e(s,a) = 0 \text{ at start, } e(s,a) \geq 0.$$

A.A. 2013-2014

6/18

<http://borghese.di.unimi.it/>



Osservazioni

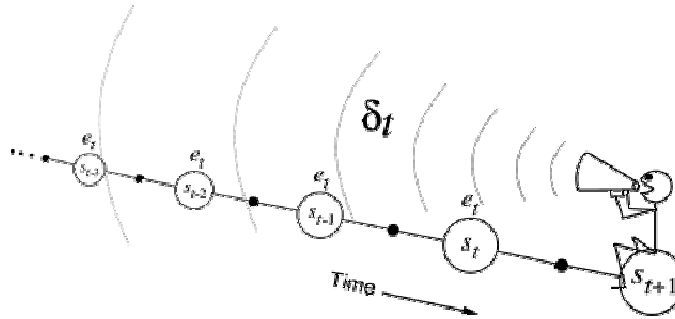


Figure 7.8: The backward or mechanistic view. Each update depends on the current TD error combined with traces of past events.

Earlier states are given less credit for the TD(0) error



Come utilizzare la eligibility trace

TD(0) Learning:

$$Q^{\pi}_{k+1}(s_t, a_t) = Q^{\pi}_k(s_t, a_t) + \alpha [r_{t+1} + \gamma Q^{\pi}_k(s_{t+1}, a_{t+1}) - Q^{\pi}_k(s_t, a_t)]$$

Errore: δ_t

$$Q^{\pi}_{k+1}(s_t, a_t) = Q^{\pi}_k(s_t, a_t) + \alpha \delta_t \quad \text{Per 1 coppia (s,a)}$$

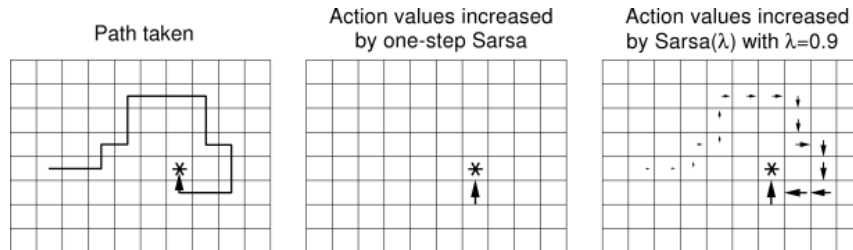
$$Q^{\pi}_{k+1}(s, a) = Q^{\pi}_k(s, a) + \alpha \delta_t e(s, a) \quad \text{Per tutte le coppie (s,a)}$$

Propagate error at time t to all (s,a)

Eleggibilità: $e_t(s,a)$



Esempio



Con il semplice costo di una variabile per ogni coppia stato-azione, ho un aggiornamento graduale della funzione valore di più stati.

$Q^\pi(s,a)$ inizializzati ad un valore leggermente negativo.
 $r = 0$ per ogni stato prossimo, tranne lo stato finale, per il quale $r = +1$.



SARSA(λ)

Initialize $Q(s, a)$ arbitrarily and $e(s, a) = 0$, for all s, a
 Repeat (for each episode):
 Initialize s, a
 Repeat (for each step of episode):
 Take action a , observe r, s'
 Choose a' from s' using policy derived from Q (e.g., ϵ -greedy)
 $\delta \leftarrow r + \gamma Q(s', a') - Q(s, a)$
 $e(s, a) \leftarrow e(s, a) + 1$
 For all s, a :
 $Q(s, a) \leftarrow Q(s, a) + \alpha \delta e(s, a)$
 $e(s, a) \leftarrow \gamma \lambda e(s, a)$
 $s \leftarrow s'; a \leftarrow a'$
 until s is terminal



Trace and Q(λ)

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right]$$

Quanto posso propagare all'indietro l'"errore"?

NB L'azione che scelgo può non essere la migliore, la funzione Q viene modificata run-time.

Watkin's version



Watkin's Q(λ)

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right]$$

Suppongo di scegliere $a' = a_{t+1}$ azione esplorativa secondo la policy π .

Posso sempre calcolare $Q(s_t, a_t)$, scegliendo il $\max(Q(s_{t+1}, a_{t+1}))$. Questo vuole dire ipotizzare di scegliere $a_{\max} = \operatorname{argmax}(\max(Q(s_{t+1}, a_{t+1})))$, che in questo caso: $a_{\max} \neq a'$.

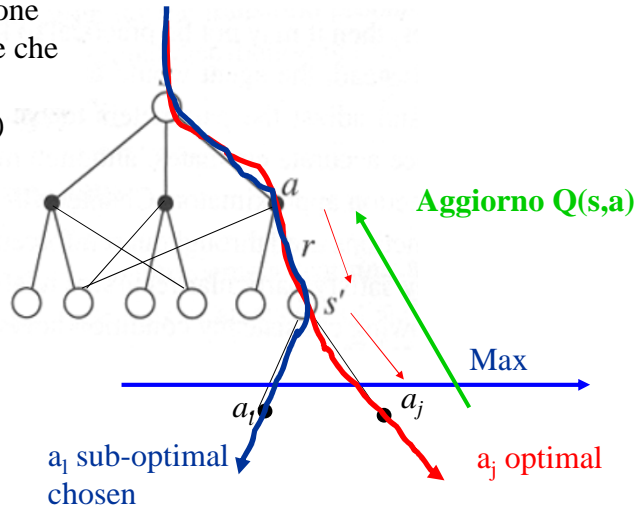
Ma poi devo ripartire da capo perchè da lì in poi seleziono una sequenza diversa di transizioni di stato.



Analisi grafica delle mosse ϵ -esplorative



Se scelgo un'azione diversa, vuol dire che $Q(s, a_i) > Q(s, a_j)$, facilmente $Q(s, a)$ aumenta.



A.A. 2013-2014

13/18

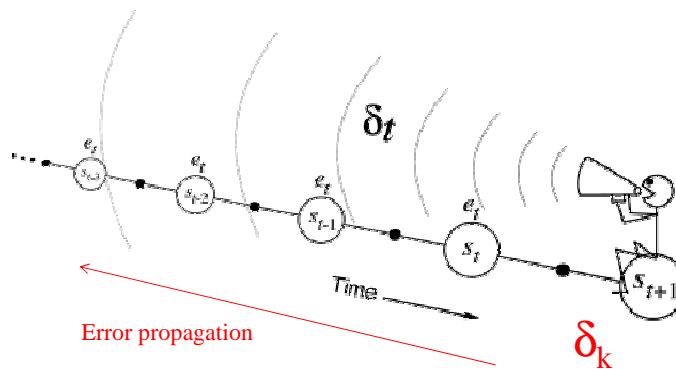
<http://borghese.di.unimi.it/>



Osservazioni



When to stop propagation?



When $s_{t-k} = s_{t+1}$

A.A. 2013-2014

14/18

<http://borghese.di.unimi.it/>



New strategy for updating eligibility trace



Eligibility set to 0 for the states in which the non-greedy action was chosen.

First decay or set to 0 the eligibility.

Then increment by 1 the eligibility of the current state.

$$e_t(s, a) = \bar{I}_{sst} \cdot \bar{I}_{aat} + \begin{cases} \gamma \lambda e_{t-1}(s, a) & \text{if } Q_{t-1}(s_t, a_t) = \max_a Q_{t-1}(s_t, a); \\ 0 & \text{otherwise,} \end{cases}$$



Q-learning



$$e_t(s, a) = \bar{I}_{sst} \cdot \bar{I}_{aat} + \begin{cases} \gamma \lambda e_{t-1}(s, a) & \text{if } Q_{t-1}(s_t, a_t) = \max_a Q_{t-1}(s_t, a); \\ 0 & \text{otherwise,} \end{cases}$$

Aggiorno Q:

$$Q_{t+1}(s, a) = Q_t(s, a) + \alpha \delta_t e_t(s, a),$$

$$\delta_t = r_{t+1} + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t).$$

Scelta di a:

Se scelgo a_{\max} , continuo come SARSA, altrimenti $e(s, a) = 0$.



Algorithm for Watkin's $Q(\lambda)$

```
Initialize  $Q(s, a)$  arbitrarily and  $e(s, a) = 0$ , for all  $s, a$ 
Repeat (for each episode):
  Initialize  $s, a$ 
  Repeat (for each step of episode):
    Take action  $a$ , observe  $r, s'$ 
    Choose  $a'$  from  $s'$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)
     $a^* \leftarrow \arg \max_b Q(s', b)$  (if  $a'$  ties for the max, then  $a^* \leftarrow a'$ )
     $\delta \leftarrow r + \gamma Q(s', a^*) - Q(s, a)$ 
     $e(s, a) \leftarrow e(s, a) + 1$ 
    For all  $s, a$ :
       $Q(s, a) \leftarrow Q(s, a) + \alpha \delta e(s, a)$ 
      If  $a' = a^*$ , then  $e(s, a) \leftarrow \gamma \lambda e(s, a)$ 
      else  $e(s, a) \leftarrow 0$ 
     $s \leftarrow s'; a \leftarrow a'$ 
  until  $s$  is terminal
```



Sommario

$Q(\lambda)$