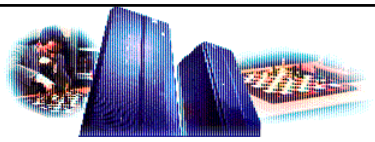
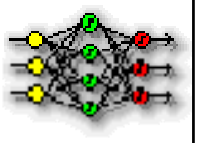




Sistemi Intelligenti
Reinforcement Learning: Q(λ)

Alberto Borghese

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

Sommario

Q(λ)

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 assignement.
structural credit assignement.

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.

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

$$\Delta Q^{\pi}(s_t,a_t) = +\alpha\,\delta_k$$

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

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 assignement).

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

Definisce se uno stato è eleggibile e "quanto" sia eleggibile, cioè che percentuale di aggiornamento meriti.

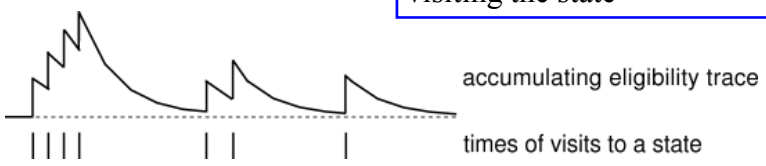
NB L'errore modifica il valore di Q(s,a) ma questo modifica a sua volta il valore di $Q(s_{t-1}, a_{t-1})$....

Eligibility trace per la funzione Q

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

decay

Increases: depends only on visiting the state



e(s,a) = 0 at start, e(s,a) ≥ 0.

Osservazioni

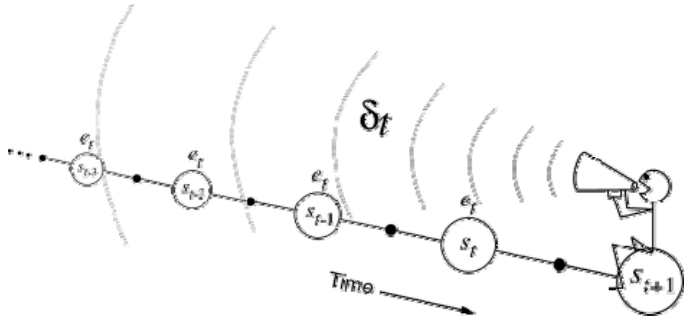


Figure 7.8: The backward or mechanistic view. Each update depends on the current TD error combined with traces of past events.

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

Come utilizzare la eligibility trace

TD(0) Learning:

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

Errore: δ_t

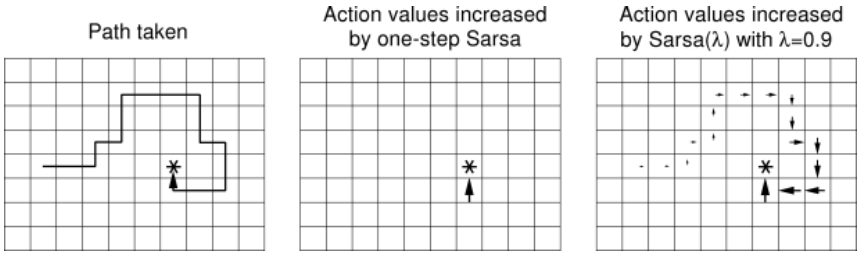
$$Q^{\pi}{}_{k+1}(s_t,a_t) = Q^{\pi}{}_k(s_t,a_t) + \alpha\delta_t$$ Per 1 coppia (s,a)

$$Q^{\pi}{}_{k+1}(s,a) = Q^{\pi}{}_k(s,a) + \alpha\delta_t e(s,a)$$ 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 policy viene modificata run-time e la funzione Q viene associata a cammini diversi nel grafo di transizione di stato.

Watkin's version

Watkin's Q(λ)

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

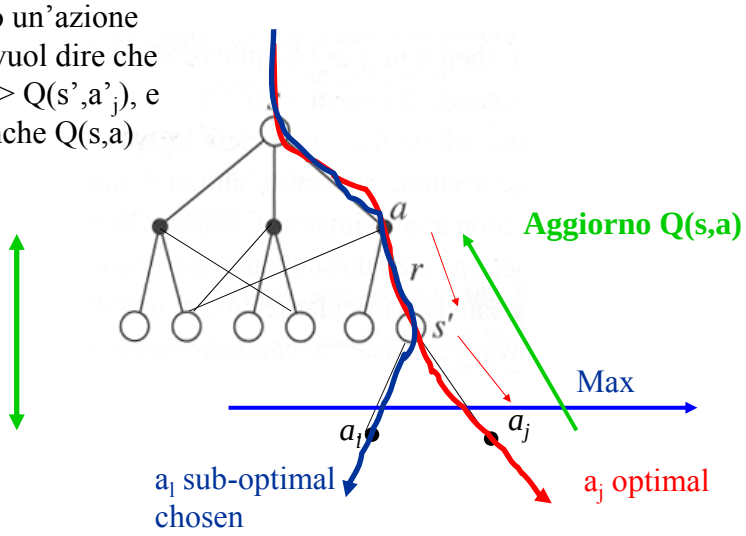
Suppongo di scegliere a' = a_{t+1} azione prescritta dalla policy π.

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} = 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. Visito il grafo in modo diverso.

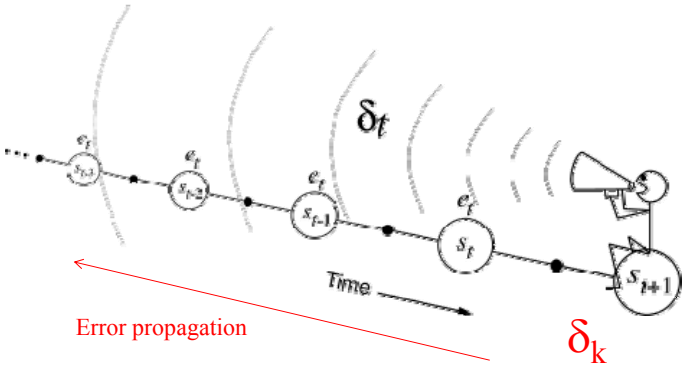
Analisi grafica delle mosse ε-eplorative

Se scelgo un'azione diversa, vuol dire che $Q(s',a'_l) > Q(s',a'_j)$, e quindi anche Q(s,a) aumenta.



Osservazioni

When to stop propagation?



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

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.

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

Q-learning

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

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(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)
    $a^* \leftarrow \arg\max_b Q(s',b)$ (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(λ)