Reinforcement Learning (INF11010)

Lecture 11: Eligibility Traces

Pavlos Andreadis, March 13th 2018

Today's Content

- n-step Return in TD Learning
- TD(λ) prediction
  - Forward view
  - Backward view
- TD(λ) control
  - Sarsa(λ)
  - Q(λ)

Previously...

- Return:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^{T-t-1} r_T$$

- 1-step TD Return:

$$R_t^{(1)} = r_{t+1} + \gamma V_t(s_{t+1})$$

Now...

- Return:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^{T-t-1} r_T$$

- 1-step TD Return:

$$R_t^{(1)} = r_{t+1} + \gamma V_t(s_{t+1})$$

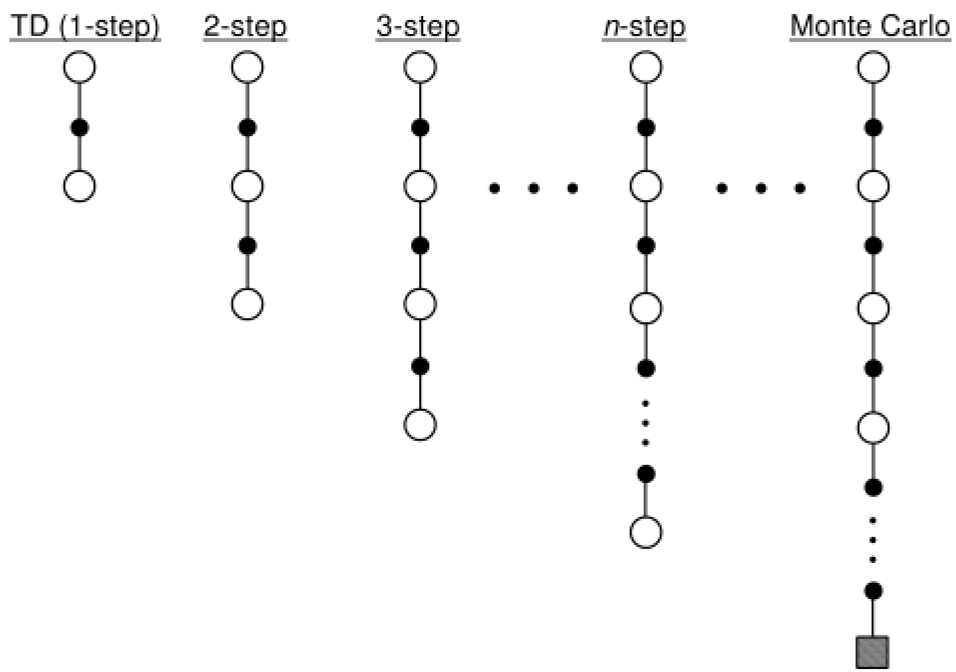
- 2-step TD Return:

$$R_t^{(2)} = r_{t+1} + \gamma r_{t+2} + \gamma^2 V_t(s_{t+2})$$

- n-step TD Return:

$$R_t^{(n)} = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n V_t(s_{t+n})$$

n-Step Return



n-Step TD policy evaluation

$$\Delta V_t(s_t) = a\left[R_t^{(n)} - V_t(s_t)\right]$$

- updates according to n rewards in the future
- will converge to correct predictions
- a theoretical tool (not particularly practical)
  - so what is used in practise? → turn page →

Forward view of TD(λ)

- complex backup example:

$$R_t^{average} = \frac{1}{2} R_t^{(2)} + \frac{1}{2} R_t^{(4)}$$

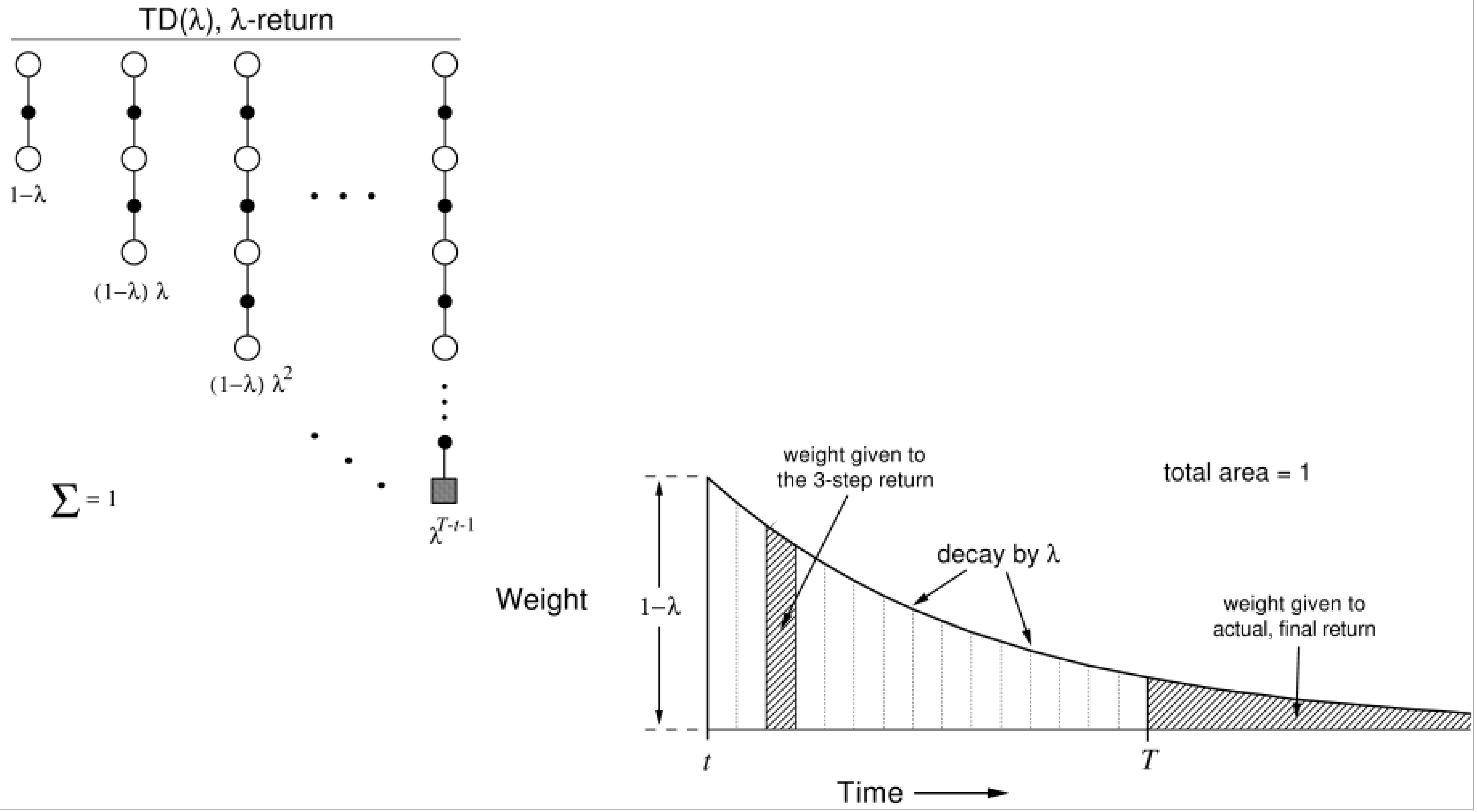
- λ-return:

$$R_t^{\lambda} = (1 - \lambda) \sum_{n=1}^{\infty} {}^{n-1} R_t^{(n)}$$

- whenever not enough steps ahead → full return
- λ-return algorithm uses the update:

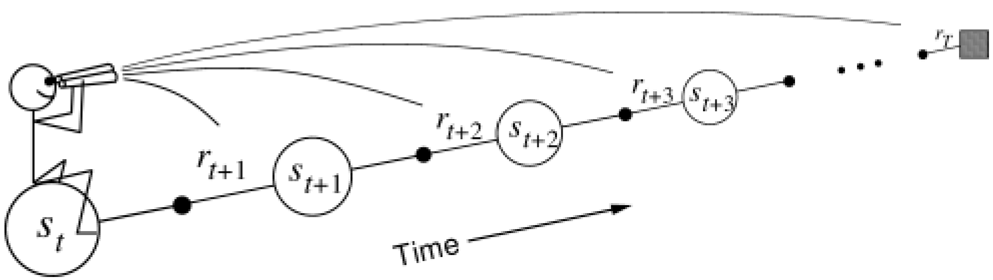
$$\Delta V_t(s_t) = a\big[R_t^{(\lambda)} - V_t(s_t)\big]$$

Forward view of TD(λ)

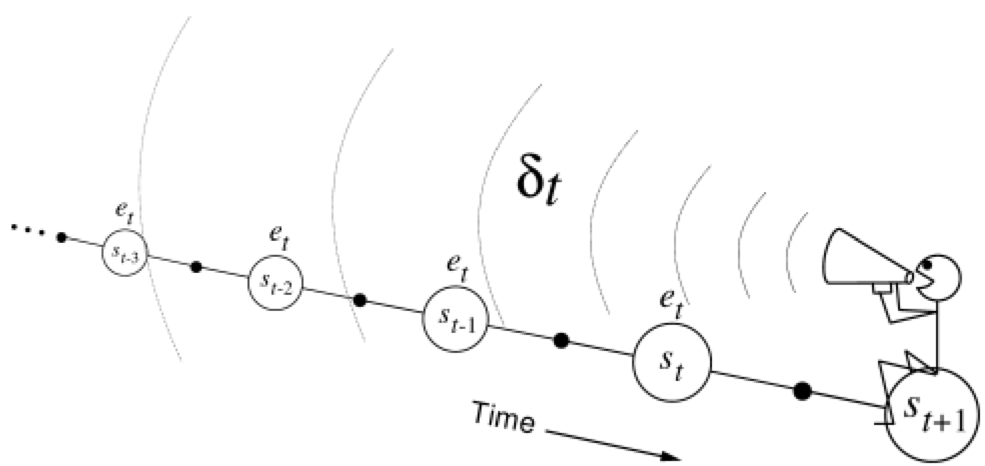


Backward view of TD(λ)

Forward view:



Backward view:



Backwards view of TD(λ)

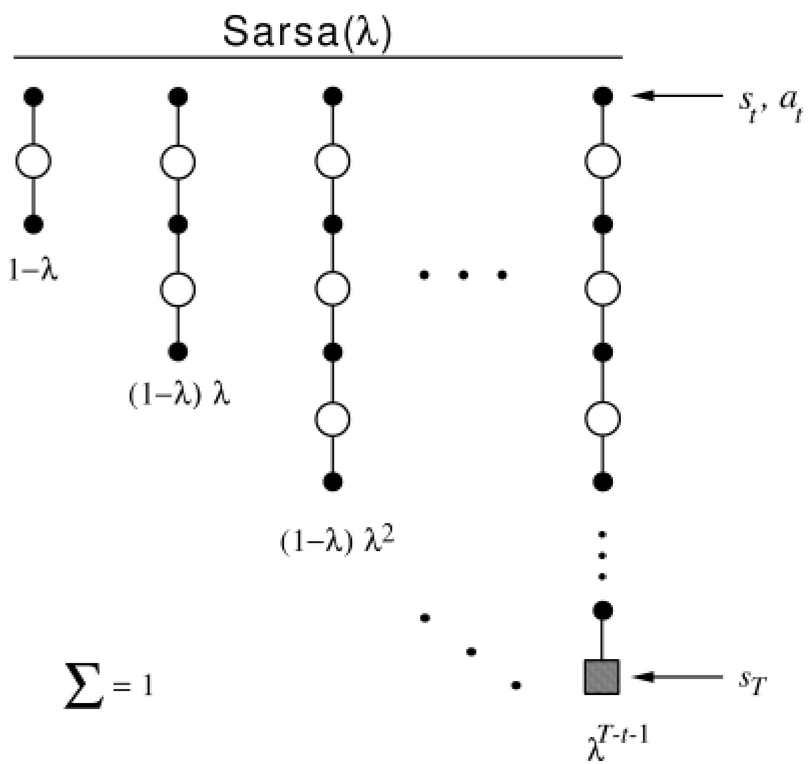
- Incremental mechanism for approximating the forward view.
- Exact for the off-line case.
- (accumulating) eligibility trace:

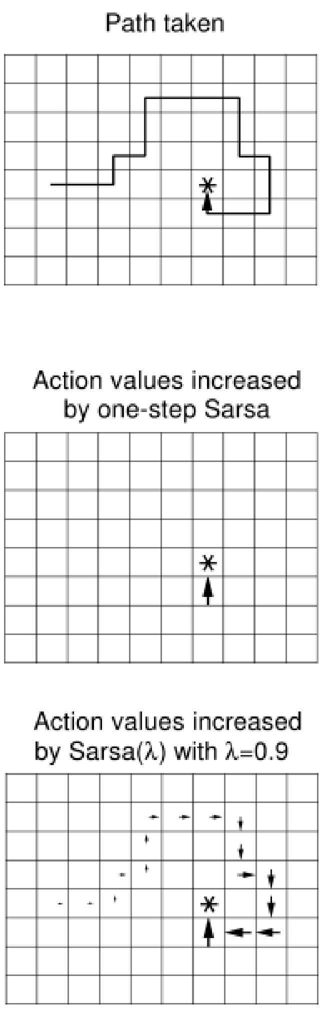
$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s) & \text{if} s \neq s_t; \\ \gamma\lambda e_{t-1}(s) + 1 & \text{if} s = s_t \end{cases}$$

- decay: $\gamma\lambda$
- trace-decay parameter: λ
- TD error: $\delta_t = r_{t+1} + \gamma V_t(s_{t+1}) - V_t(s_t)$
- TD(λ) update:

$$\Delta V_t(s_t) = a\delta_t e_t(s),\ \forall s \in S$$

TD(λ) control - Sarsa(λ)

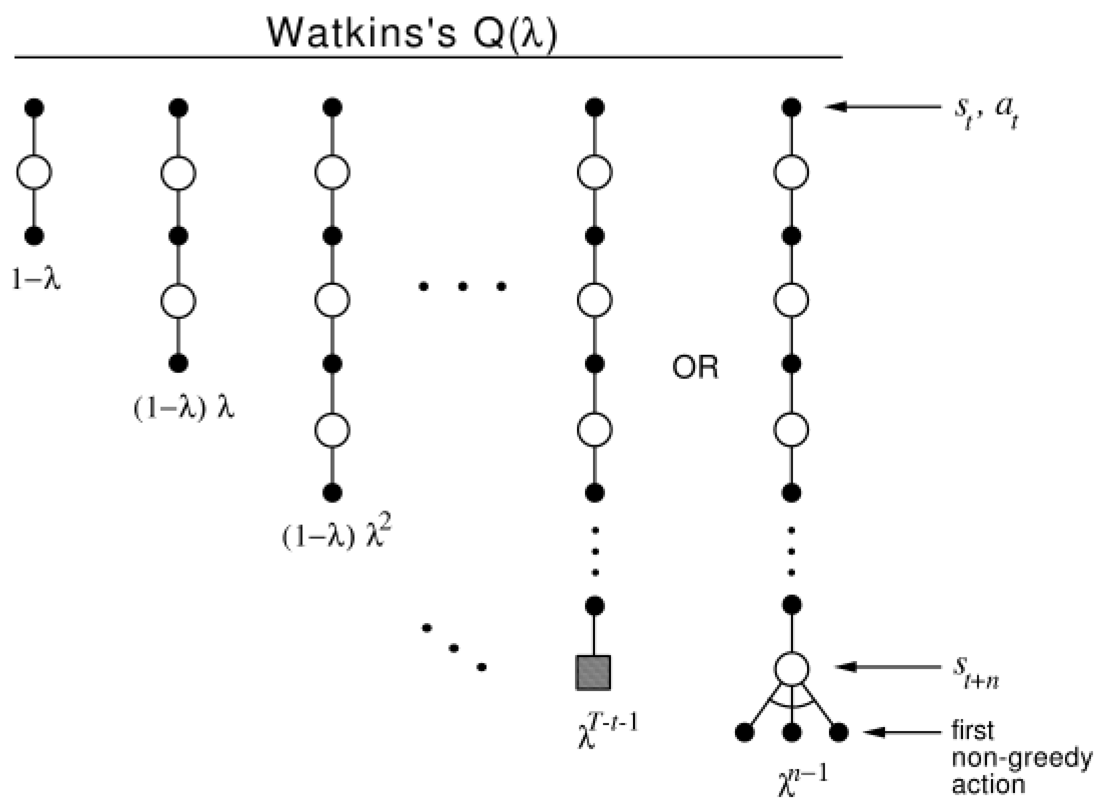




TD(λ) control - Sarsa(λ)

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

TD(λ) control - Q(λ)



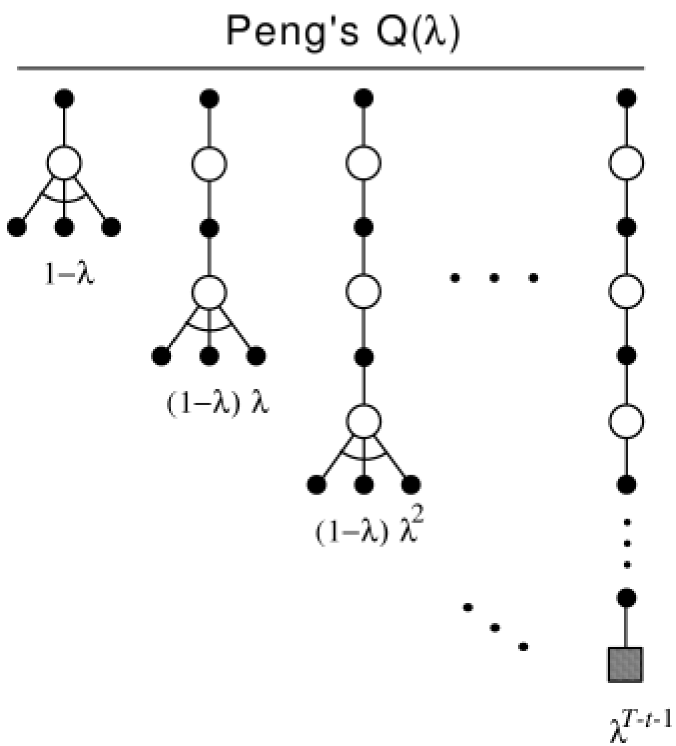
TD(λ) control - Q(λ)

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

TD(λ) control - 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

TD(λ) control - Q(λ)



Summary

- n-step Return in TD Learning
- TD(λ) prediction
  - Forward view
  - Backward view
- TD(λ) control
  - Sarsa(λ)
  - Q(λ)

Reading +

- Chapter 7 (7.1 to 7.3, 7.5 to 7.6, 7.9, 7.11) of Sutton and Barto (1st Edition) http://incompleteideas.net/book/ebook/the-book.html

<u>Optional</u>:

- Chapter 7 (the rest) of Sutton and Barto (1st Edition)