Retrace(λ)

Temporal Credit Assignment in Off-Policy Reinforcement Learning

Matteo Papini

28th November 2017

Before we start...

Syllabus

Reinforcement Learning

Value Function

Policy Iteration

Temporal Difference Learning

Value Function update at time t:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[G_t - Q(s_t, a_t) \right]$$

$$\Delta Q(s_t, a_t) = \alpha \left[G_t - Q(s_t, a_t) \right]$$

Expected Sarsa

- **Forward view**: look one step forward to compute the target

$$G_t = r_{t+1} + \gamma \underset{a \sim \pi}{E} \left[Q(s_{t+1}, a) \right]$$

- **Backward view**: wait one step to update $Q(s_t, a_t)$

$$\Delta Q(s_{t-1}, a_{t-1}) = \alpha \left[r_t + \gamma \underset{a \sim \pi}{E} \left[Q(s_t, a) \right] - Q(s_{t-1}, a_{t-1}) \right]$$

Temporal Credit Assignment Problem

n-step Expected Sarsa

Look far (n steps) in the future:

$$G_t^{(n)} \doteq r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n \mathop{E}_{a\sim\pi} [Q(s_{t+n}, a)]$$
$$= \sum_{k=1}^{n} \gamma^{k-1} r_{t+k} + \gamma^n \mathop{E}_{a\sim\pi} [Q(s_{t+n}, a)]$$

- $G_t^{(1)}$ is a TD target
- $G_t^{(T-t)}$ is a Monte Carlo target

Forward View

Average **all** n-step targets:

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

- $\lambda = 0$ gives $G_t^{(1)}$, the TD target
- $\lambda = 1$ gives $G_t^{(T-t)}$, the Monte Carlo target

Forward View

Average **all** n-step targets:

$$G_t^{\lambda} \doteq (1-\lambda)\sum_{n=1}^{T} \lambda^{n-1} G_t^{(n)}$$

- $\lambda = 0$ gives $G_t^{(1)}$, the TD target
- $\lambda = 1$ gives $G_t^{(T-t)}$, the Monte Carlo target

Forward View

Average **all** n-step targets:

$$G_t^{\lambda} = (1-\lambda)\sum_{n=1}^{T-t-1} \lambda^{n-1} G_t^{(n)} + \lambda^{T-t-1} G_t^{(T-t)}$$

- $\lambda = 0$ gives $G_t^{(1)}$, the TD target
- $\lambda = 1$ gives $G_t^{(T-t)}$, the Monte Carlo target

Forward View

Average **all** n-step targets:

$$G_t^{\lambda} = (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} G_t^{(n)} + \lambda^{T-t-1} G_t^{(T-t)}$$

- $\lambda = 0$ gives $G_t^{(1)}$, the TD target
- $\lambda = 1$ gives $G_t^{(T-t)}$, the Monte Carlo target

Forward View

Average **all** n-step targets:

$$G_t^{\lambda} = (1-\lambda)\sum_{n=1}^{T-t-1}\lambda^{n-1}G_t^{(n)} + \lambda^{T-t-1}G_t^{(T-t)}$$

- $\lambda = 0$ gives $G_t^{(1)}$, the TD target
- $\lambda = 1$ gives $G_t^{(T-t)}$, the Monte Carlo target

Backward View

Video here

Eligibility Traces

Accumulating Traces

Update at step t:

$$e(s,a) \leftarrow \gamma\lambda e(s,a) + \mathbb{1}\{s = s_t, a = a_t\}$$
$$\Delta Q(s,a) = \alpha e(s,a)(G_t^{(1)} - Q(s,a))$$

Accumulating Traces

Update at step t:

$$e \leftarrow \gamma\lambda e + \mathbb{1}_t$$
$$\Delta Q = \alpha e \delta_t$$

Replacing Traces

Update at step t:

$$\begin{aligned} e &\leftarrow (1 - \mathbb{1}_t)\gamma\lambda e + \mathbb{1}_t \\ \Delta Q &= \alpha e \delta_t \end{aligned}$$

Dutch Traces

Update at step t:

$$e \leftarrow (1 - \alpha \mathbb{1}_t)\gamma\lambda e + \mathbb{1}_t$$
$$\Delta Q = \alpha e \delta_t$$

Update at time t:

$$\Delta Q = \alpha e_t \delta_t$$

General Traces

Update at time t:

$$\Delta Q = \alpha\gamma^t \left(\prod_{s=1}^{t} c_s\right) \delta_t$$

Title B1

- Lorem ipsum dolor sit amet, consectetur adipiscing elit.
- Nulla id ex ornare, gravida nisi in, ornare risus.
    1. Aenean eu posuere purus.
    2. Etiam maximus convallis libero, ac venenatis nunc sagittis nec.
- Suspendisse orci ex, pharetra vitae aliquam ac, rutrum in dui.

Theorem (Th. Name)

This is a theorem

- *Property 1;*
- *Property 2.*

Proof.

$$a + b = c \quad (1)$$

$$a = c - b \quad (2)$$

$$answer = 42 \quad (3)$$

□

Proof.

Another proof style. □

Title B2

Theorem (Th. Name)

This is a theorem

- *Property 1;*
- *Property 2.*

Proof.

$$a + b = c \tag{1}$$

$$a = c - b \tag{2}$$

$$answer = 42 \tag{3}$$

□

Proof.

Another proof style. □

Theorem (Th. Name)

This is a theorem

- *Property 1;*
- *Property 2.*

Proof.

$$a + b = c \tag{1}$$
$$a = c - b \tag{2}$$
$$answer = 42 \tag{3}$$

□

Proof.

Another proof style. □

Title C1

First column.

Appears with third column

Second column.

Third column.

Title C1

First column.

Appears with third column

Second column.

Third column.

Title C1

First column.

Appears with third column

Second column.

Third column.

Title C2

Image:

Title C3

1. lorem
2. Ipsus
    1. sub1
    2. sub3
        1. sub4
        2. sub5

Video