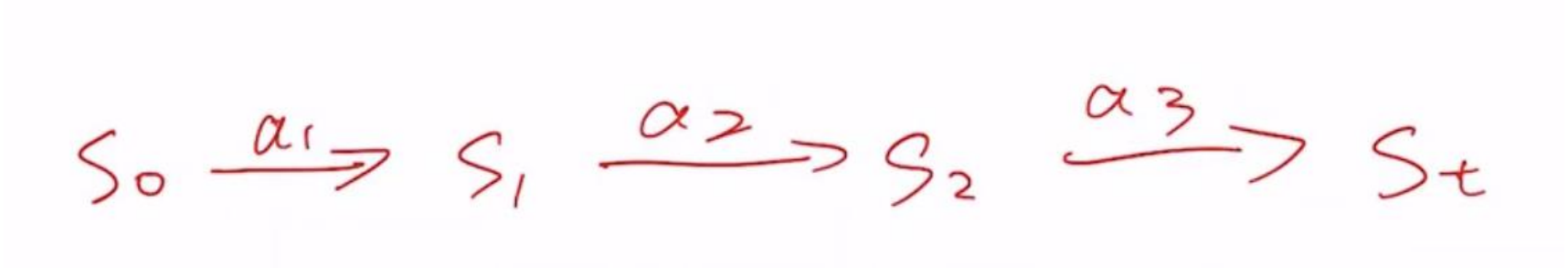
Workshop 8

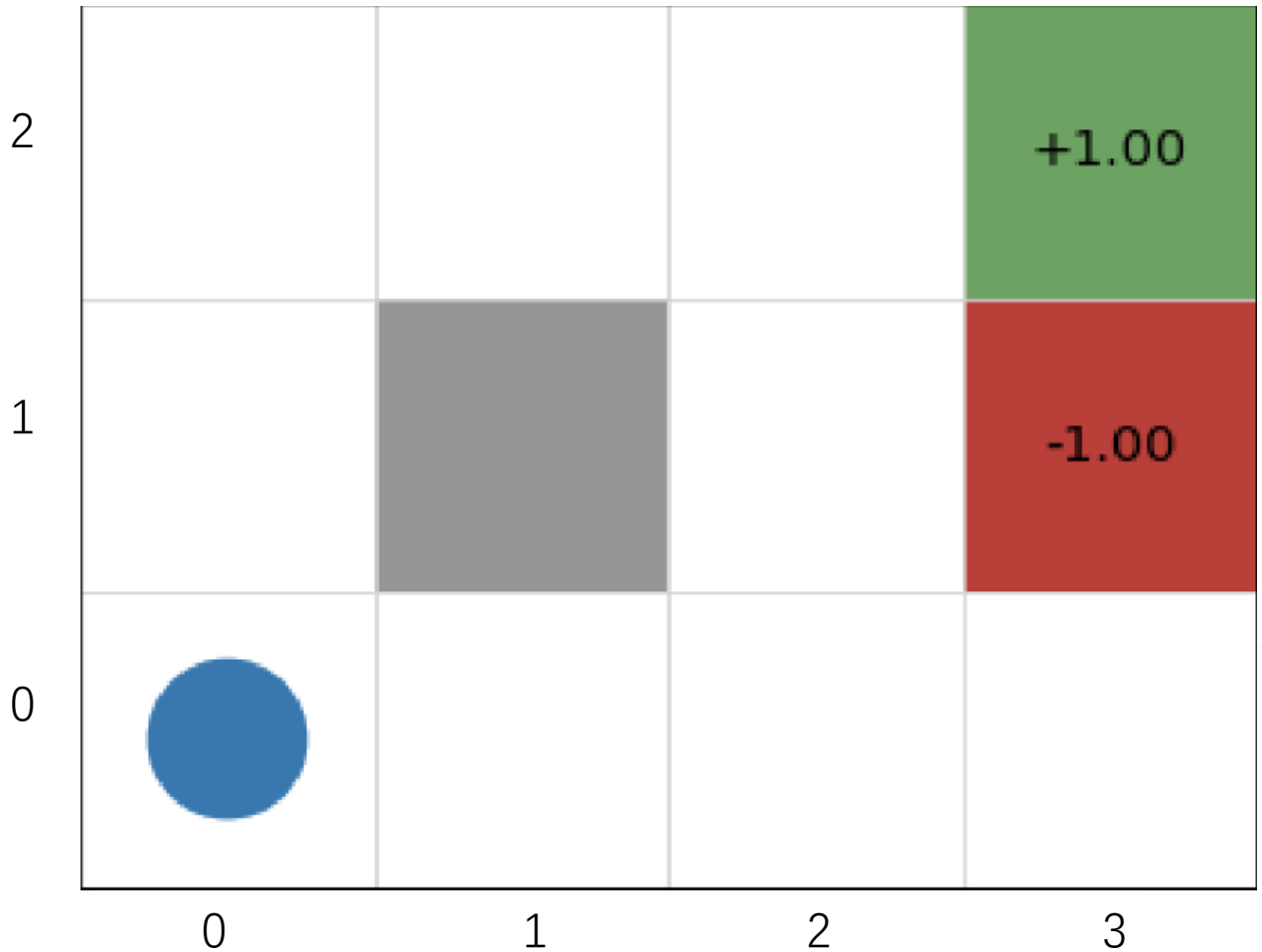
Temporal difference learning

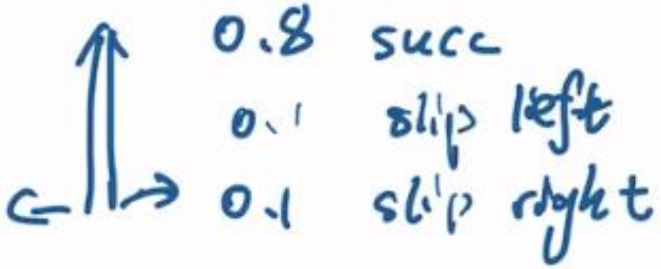


$$\delta \leftarrow [\overbrace{r}^{\text{reward}} + \overbrace{\gamma}^{\text{discount factor}} \cdot \overbrace{\max_{a'} Q(s', a')}^{V(s') \text{ estimate}}]_{\text{TD target}} - \overbrace{Q(s, a)}^{\text{do not count extra } Q(s,a)}$$

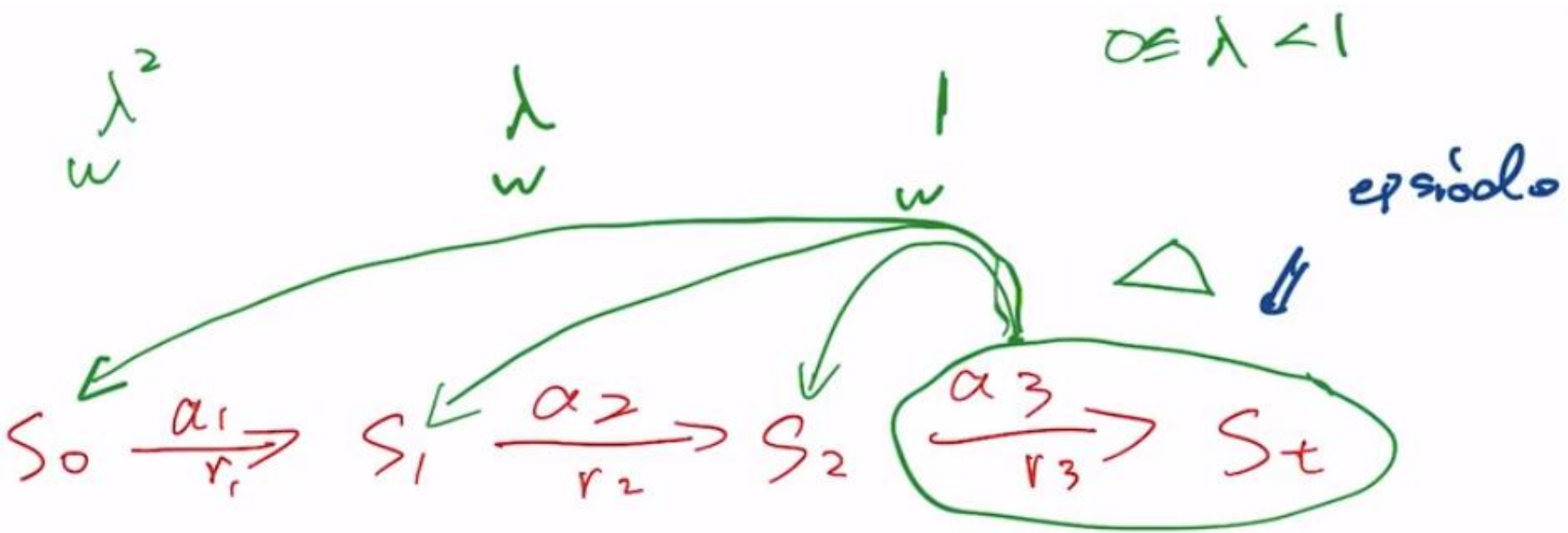
$$Q(s, a) \leftarrow \underbrace{Q(s, a)}_{\text{old value}} + \overbrace{\alpha}^{\text{learning rate}} \cdot \underbrace{\delta}_{\text{delta value}}$$

Workshop Example

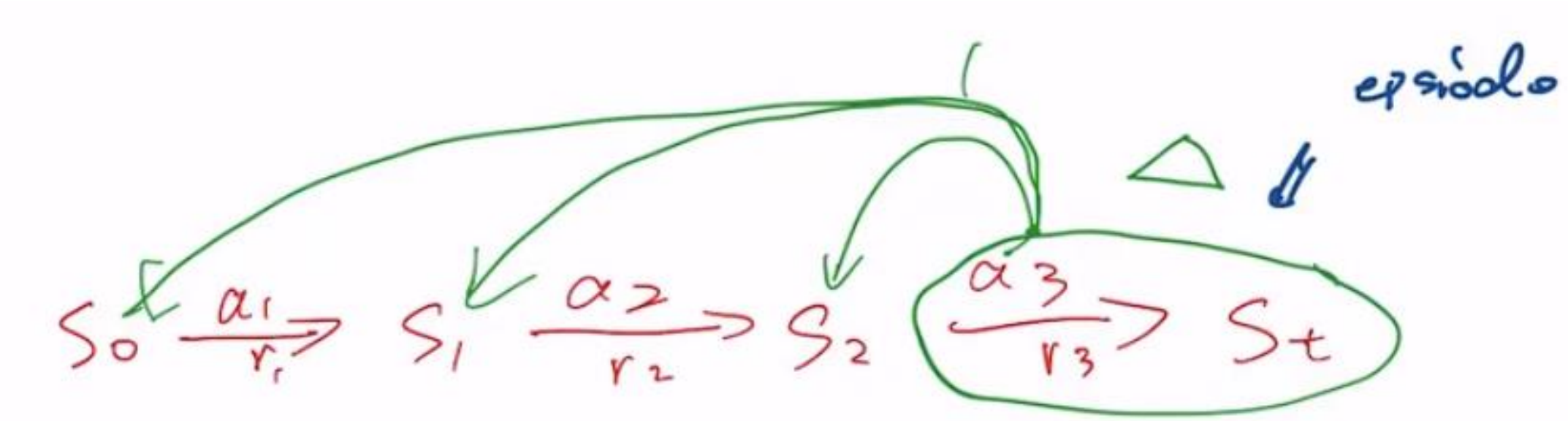




How to update Q(s, a)? TD-lambda



N-step TD



Q-learning

Repeat (for each episode)

$s \leftarrow$ the first state in episode e

Repeat (for each step in episode e)

Select action a to apply in s; e.g. Q and a multi-armed bandit algorithm

Execute action a in state s

Observe reward r and new state s'

$\delta \leftarrow r + \gamma \cdot \max_{a'} Q(s', a') - Q(s, a)$

$Q(s, a) \leftarrow Q(s, a) + \alpha \cdot \delta$

$s \leftarrow s'$

Until s is the last state of episode e (a terminal state)

Sarsa

Repeat (for each episode)

$s \leftarrow$ the first state in episode e

Select action a to apply in s using Q and a multi-armed bandit algorithm

Repeat (for each step in episode e)

Execute action a in state s

Observe reward r and new state s'

Select action a' to apply in s' using Q and a multi-armed bandit algorithm

$\delta \leftarrow r + \gamma \cdot Q(s', a') - Q(s, a)$

$Q(s, a) \leftarrow Q(s, a) + \alpha \cdot \delta$

$s \leftarrow s'$

$a \leftarrow a'$

Until s is the last state of episode e (a terminal state)

Original Q(s, a) Updating function

$$\delta \leftarrow [\overbrace{r}^{\text{reward}} + \overbrace{\gamma}^{\text{discount factor}} \cdot \overbrace{\max_{a'} Q(s', a')}^{V(s') \text{ estimate}} \overbrace{-Q(s, a)}^{\text{do not count extra } Q(s,a)}]$$

$$\text{TD target}$$

$$Q(s, a) \leftarrow \underbrace{Q(s, a)}_{\text{old value}} + \overbrace{\alpha}^{\text{learning rate}} \cdot \underbrace{\delta}_{\text{delta value}}$$

How delta differ for Q-learning and Sarsa

$$Q(s,a) \leftarrow \underbrace{Q(s,a)}_{\text{old value}} + \overbrace{\alpha}^{\text{learning rate}} \cdot \underbrace{\delta}_{\text{delta value}}$$

Q-learning

$$\delta \leftarrow r + \gamma \cdot \max_{a'} Q(s',a') - Q(s,a)$$

Sarsa

$$\delta \leftarrow r + \gamma \cdot Q(s',a') - Q(s,a)$$

Question 2

$$Q(s,a) \leftarrow \underbrace{Q(s,a)}_{\text{old value}} + \overbrace{\alpha}^{\text{learning rate}} \cdot \underbrace{\delta}_{\text{delta value}}$$

$$\delta \leftarrow r + \gamma \cdot \max_{a'} Q(s',a') - Q(s,a)$$

$$\begin{aligned} Q(S,P) &= Q(S,P) + 0.4 \cdot [r(S,P) + 0.9 \cdot \max_{a' \in A(M)} Q(M,a') - Q(S,P)] \\ &= -0.7 + 0.4 \cdot [(-1) + 0.9 \cdot (-0.4) - (-0.7)] \\ &= -0.7 + 0.4 \times (-0.66) \\ &= -0.964 \end{aligned}$$

Question 2

$$Q(s,a) \leftarrow \underbrace{Q(s,a)}_{\text{old value}} + \overbrace{\alpha}^{\text{learning rate}} \cdot \underbrace{\delta}_{\text{delta value}}$$

$$\delta \leftarrow r + \gamma \cdot \max_{a'} Q(s',a') - Q(s,a)$$

$$\begin{aligned} Q(S,P) &= Q(S,P) + 0.4 \cdot [r(S,P) + 0.9 \cdot \max_{a' \in A(M)} Q(M,a') - Q(S,P)] \\ &= -0.7 + 0.4 \cdot [(-1) + 0.9 \cdot (-0.4) - (-0.7)] \\ &= -0.7 + 0.4 \times (-0.66) \\ &= -0.964 \end{aligned}$$

Question 3

$$Q(s,a) \leftarrow \underbrace{Q(s,a)}_{\text{old value}} + \overbrace{\alpha}^{\text{learning rate}} \cdot \underbrace{\delta}_{\text{delta value}}$$

$$\delta \leftarrow r + \gamma \cdot Q(s', a') - Q(s, a)$$

$$\begin{aligned} Q(S,P) &= Q(S,P) + 0.4 \cdot [r(S,P) + 0.9 \cdot Q(M, \pi(M)) - Q(S,P)] \\ &= -0.7 + 0.4 \cdot [(-1) + 0.9 \cdot (-0.8) - (-0.7)] \\ &= -0.7 + 0.4 \times (-1.102) \\ &= -1.108 \end{aligned}$$

Question 4

$$Q(s,a) = Q(s,a) + \alpha[G_t^n - Q(s,a)]$$

$$G_t^n = r_t + \gamma \cdot t_{t+1} + \gamma^2 \cdot r_{t+2} + \cdots + \gamma^n \cdot Q(S_{t+n}, \pi(S_{t+n}))$$

$$\begin{aligned} Q(S,P) &= Q(S,P) + 0.4 \cdot [G_S^3 - Q(S,P)] \\ &= -0.7 + 0.4 \cdot [-1.4716 - (-0.7)] \\ &= -0.7 + 0.4 \times (-0.7716) \\ &= -1.00864 \end{aligned}$$

where

$$\begin{aligned} G_S^3 &= r(S,P) + 0.9 \cdot r(M,S) + 0.9^2 \cdot r(Scored,R) + 0.9^3 \cdot Q(M,P) \\ &= (-1) + 0.9 \cdot (-2) + 0.9^2 \cdot 2 + 0.9^3 \cdot (-0.4) \\ &= (-1) + (-1.8) + 1.62 + (-0.2916) \\ &= -1.4716 \end{aligned}$$