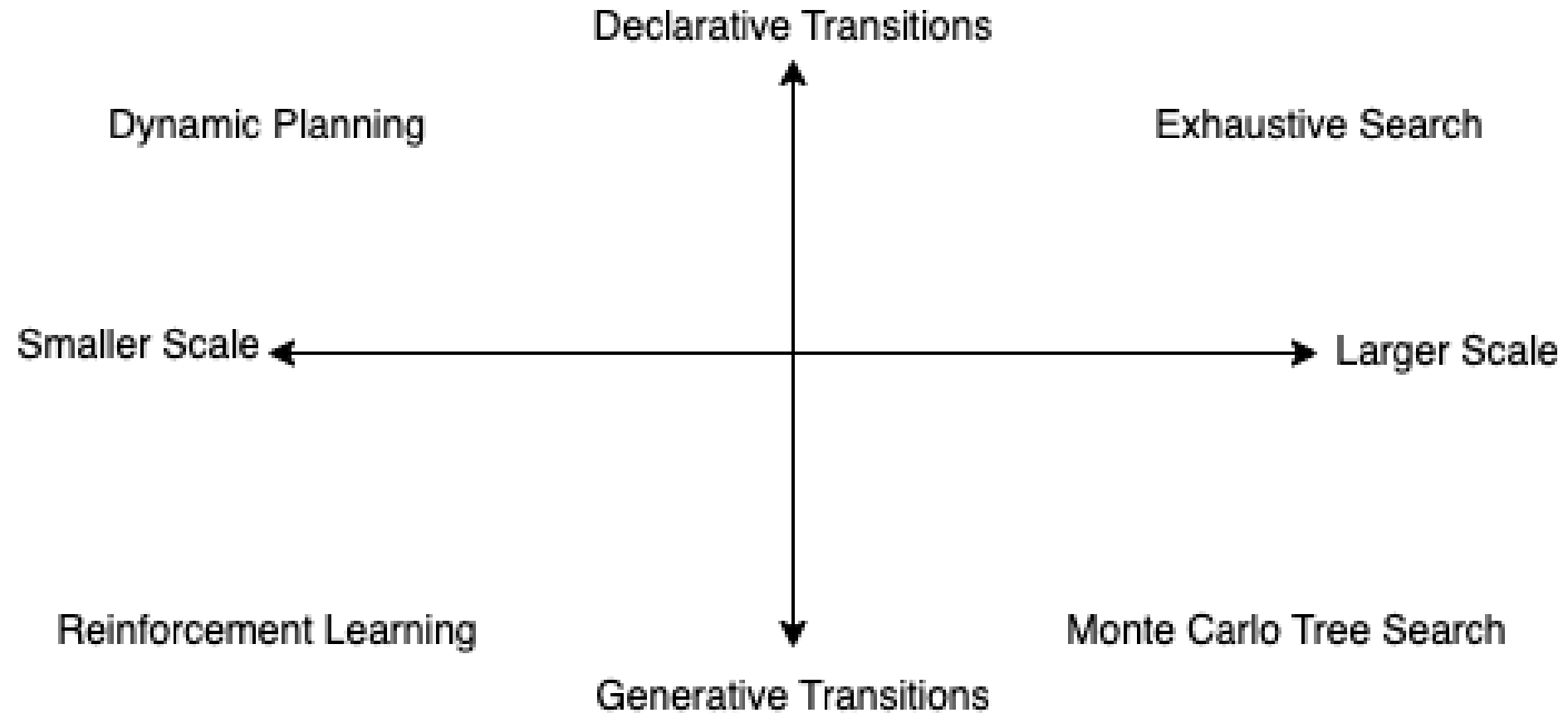


Workshop 8

Big Picture



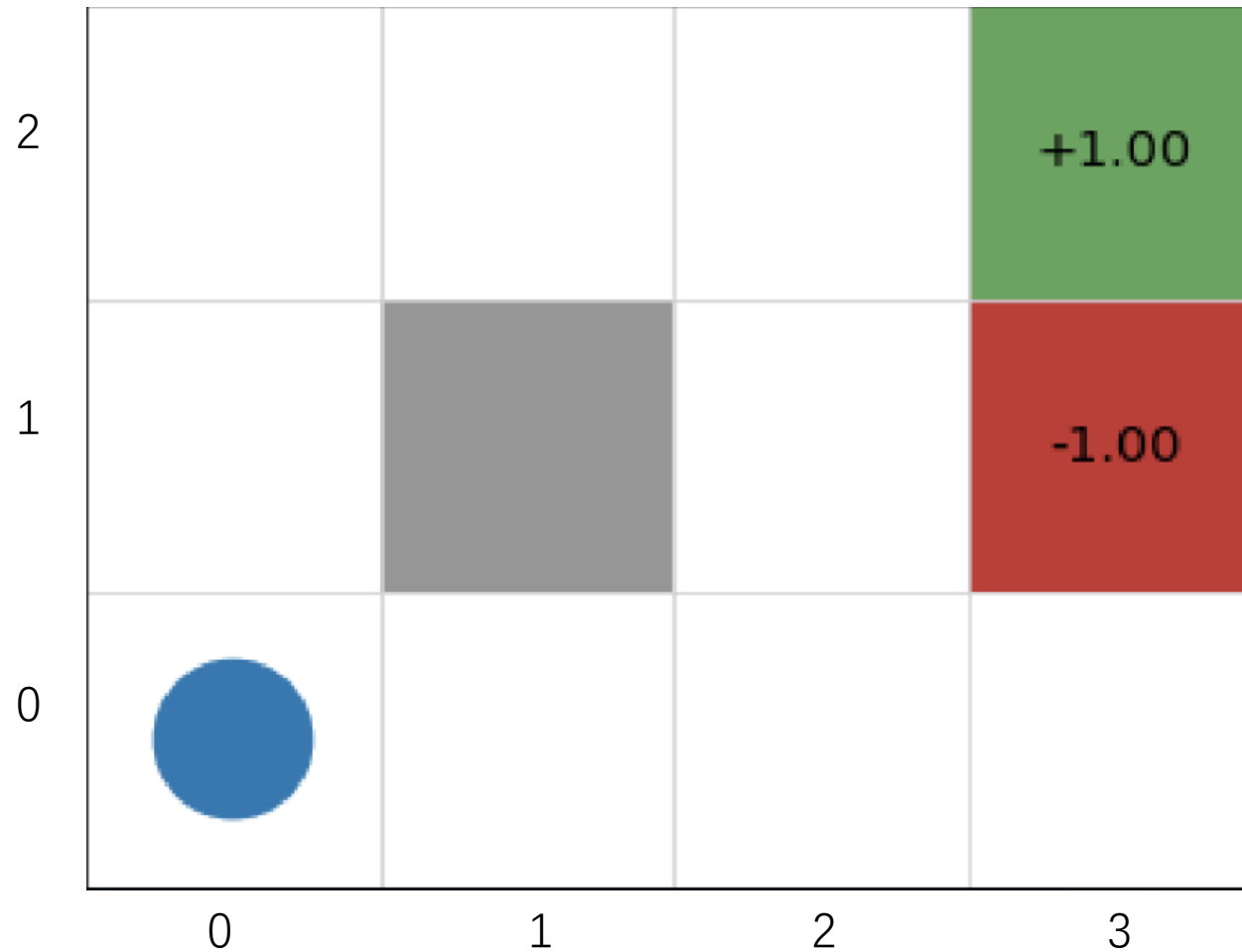
Temporal difference learning

$$s_0 \xrightarrow{\alpha_1} s_1 \xrightarrow{\alpha_2} s_2 \xrightarrow{\alpha_3} s_t$$

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

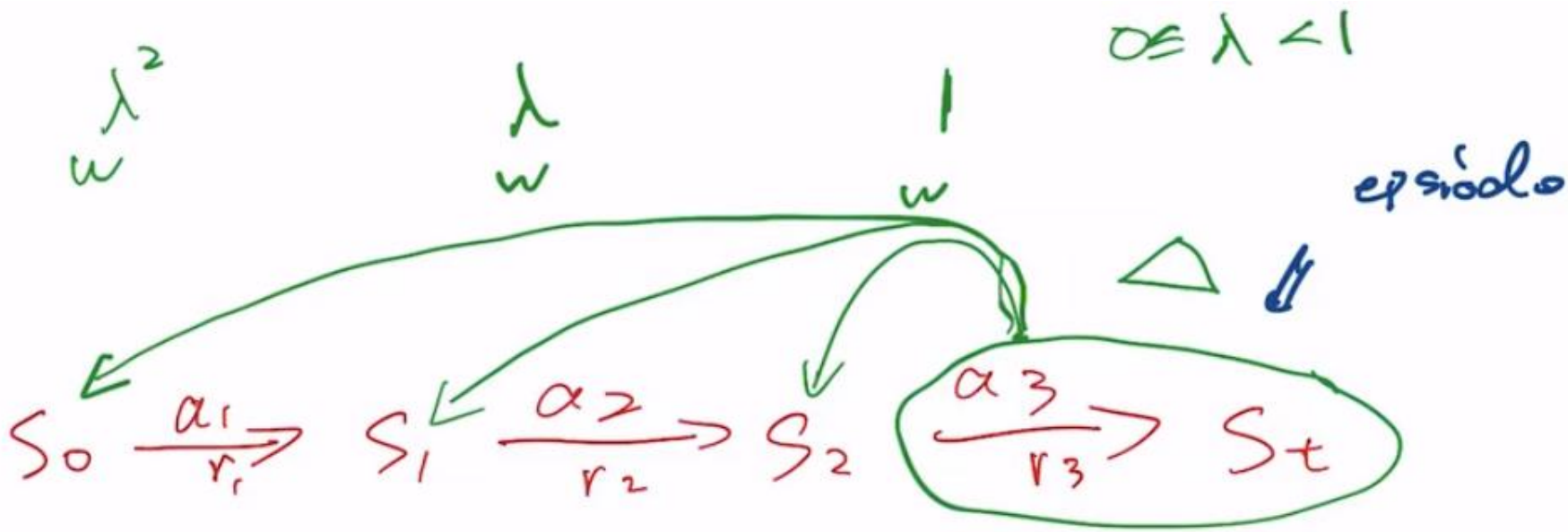
$$Q(s, a) \leftarrow \underbrace{Q(s, a)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \underbrace{\delta}_{\text{delta value}}$$

Lecture Example

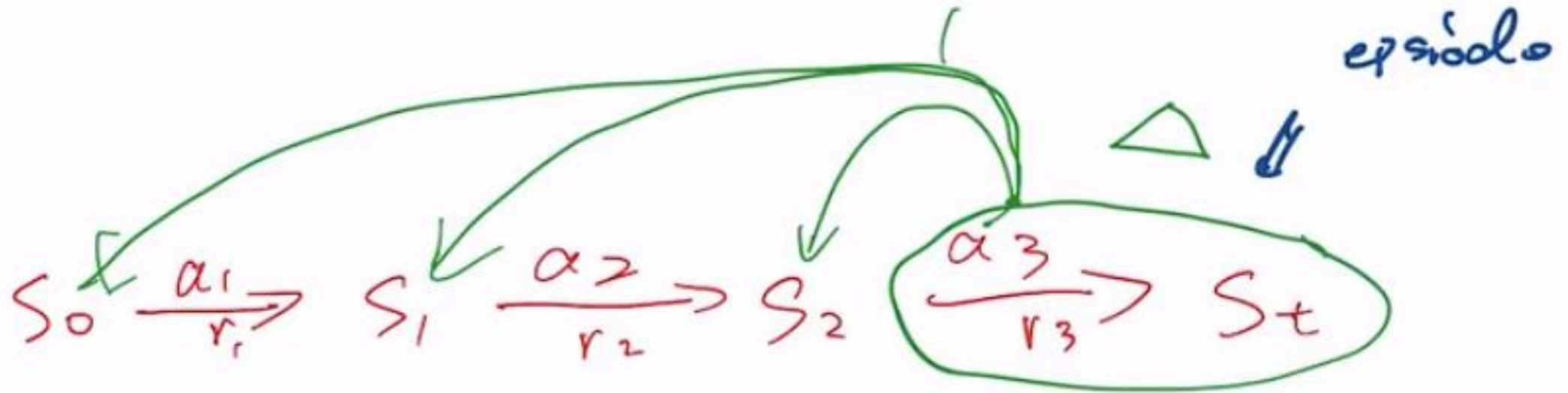


0.8 succ
0.1 slip left
0.1 slip right

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

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

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

How delta differ for Q-learning and Sarsa

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

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

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

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

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

Question 2

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

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

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

Question 3

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

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

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