Lecture 19 - March 28, 2023

- Temporal Difference Learning
  - SARSA Q-Learning
  - On-Policy Vs. off-policy
  - Expected SARSA Double Q-Learning
  - Multi-step Bootstrapping
  - SARSA-Lambda
  - Actor-Critic Method
- Function Approximation in Reinforcement Learning
  - Basics of Function Approximations
  - Least square Policy Iteration (LSPI)
  - Neural Fitted Q-Iterations (NFQI)

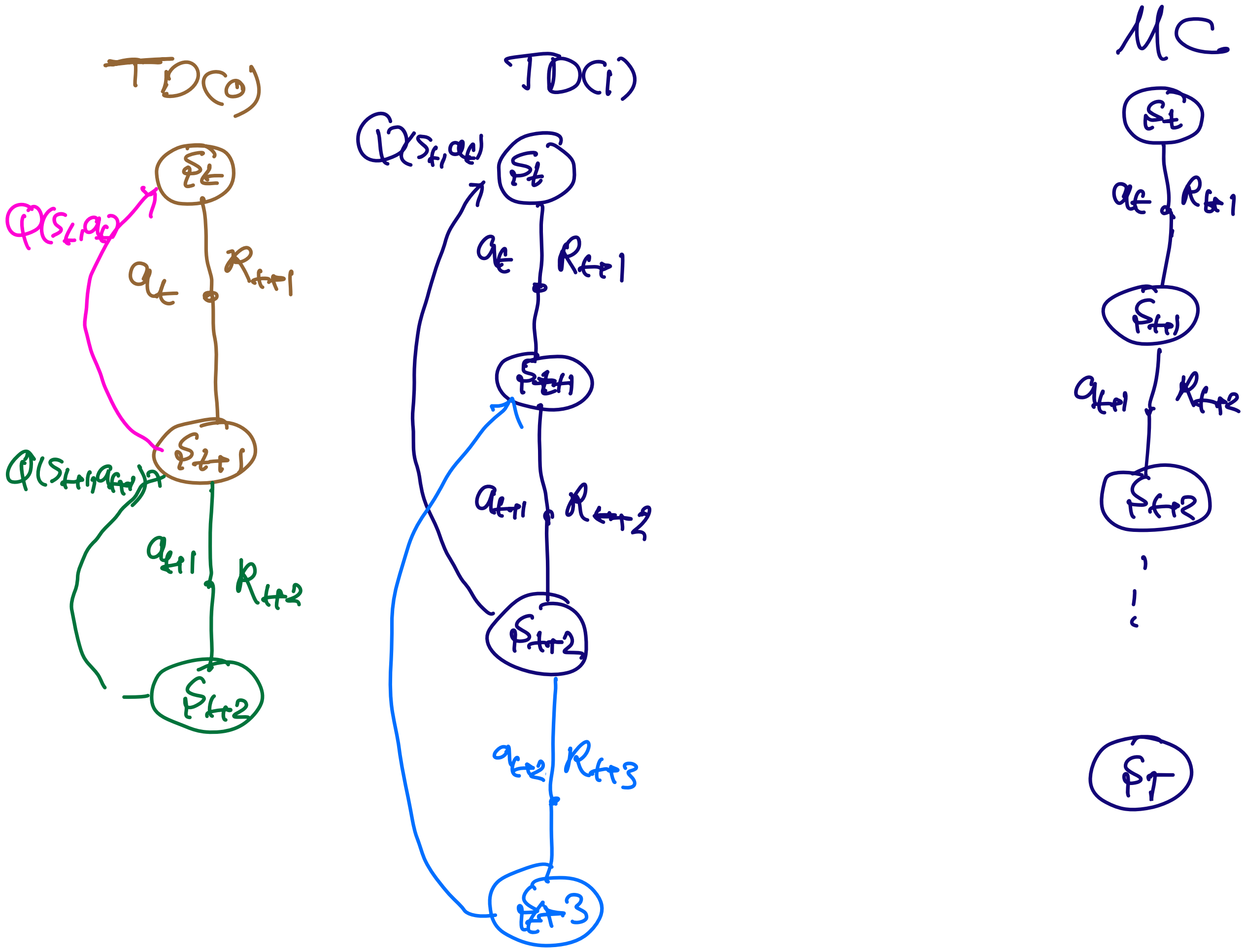
HW4 → Due March 31

Exam 2 → Tues, April 4

Project 3 → Due April 14

TA's office hour: Wendsdays, 2pm-3pm (in-person)
Fridays, 2pm-3pm (virtual)

Multi-step Bootstrapping



TD(0):

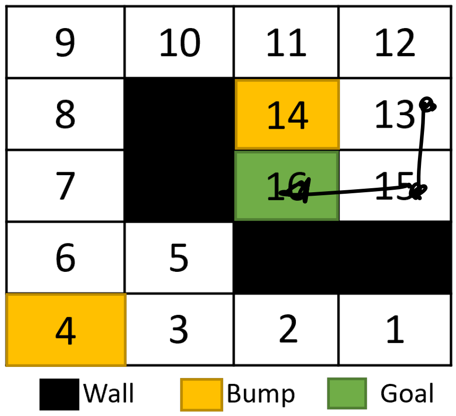
$$Q(S_t,a_t) = Q(S_t,a_t) + \alpha\left[R_{t+1} + \gamma \max Q(S_{t+1},a) - Q(S_t,a_t)\right]$$

TD(1)

$$Q(S_t,a_t) = Q(S_t,a_t) + \alpha\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 \max_a Q(S_{t+2},a) - Q(S_t,a_t)\right]$$

TD(n) larger variance $\Leftarrow$ Variance

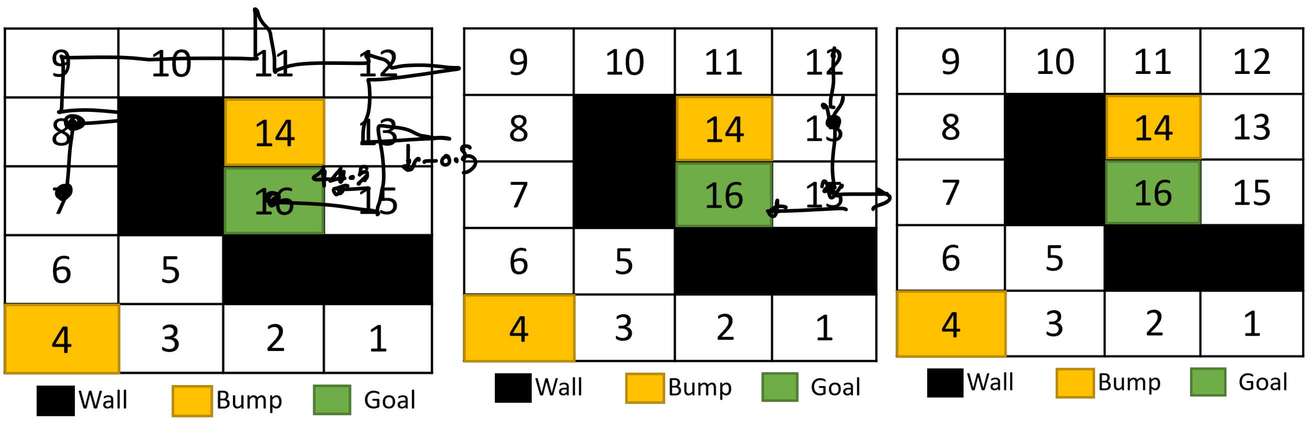
TD(n) less Biased $\Leftarrow$ Biased
than $n>0$



TD(1)

$$Q(13,D) = Q(13,D) + \alpha\left[\overset{R_{t+1}}{-1} + \gamma(99) + \gamma^2 Q(16,a) - Q(13,D)\right]$$

SARSA-Lambda — SARSA-λ



$$Q(7,U) = \underbrace{Q(7,U)}_{0} + \underbrace{\alpha}_{0.5}\left[-1 + \gamma \underbrace{Q(8,R)}_{0} - \underbrace{Q(7,U)}_{0}\right] = -0.5$$

Eligibility Trace $e(s,a)$

Larger → More Visits

Smaller → older visit

0 → No Visit

SARSA

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

SARSA Error δ_t

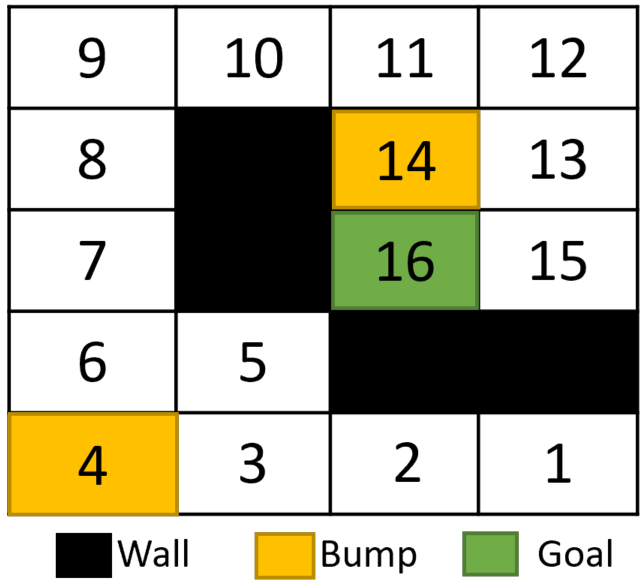
observe → $e(s_t,a_t) = e(s_t,a_t) + 1$

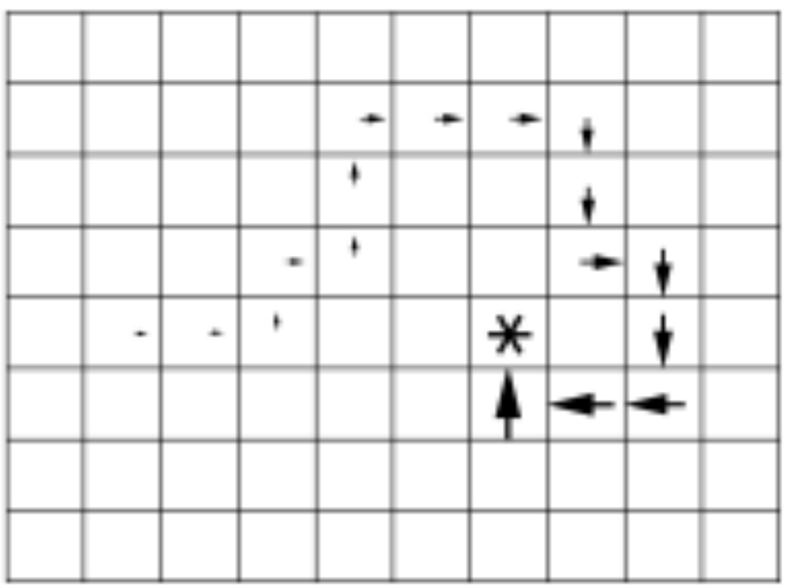
$$Q(s,a) = Q(s,a) + \alpha\, \delta_t\, e(s,a) \quad \text{for all } s, a$$

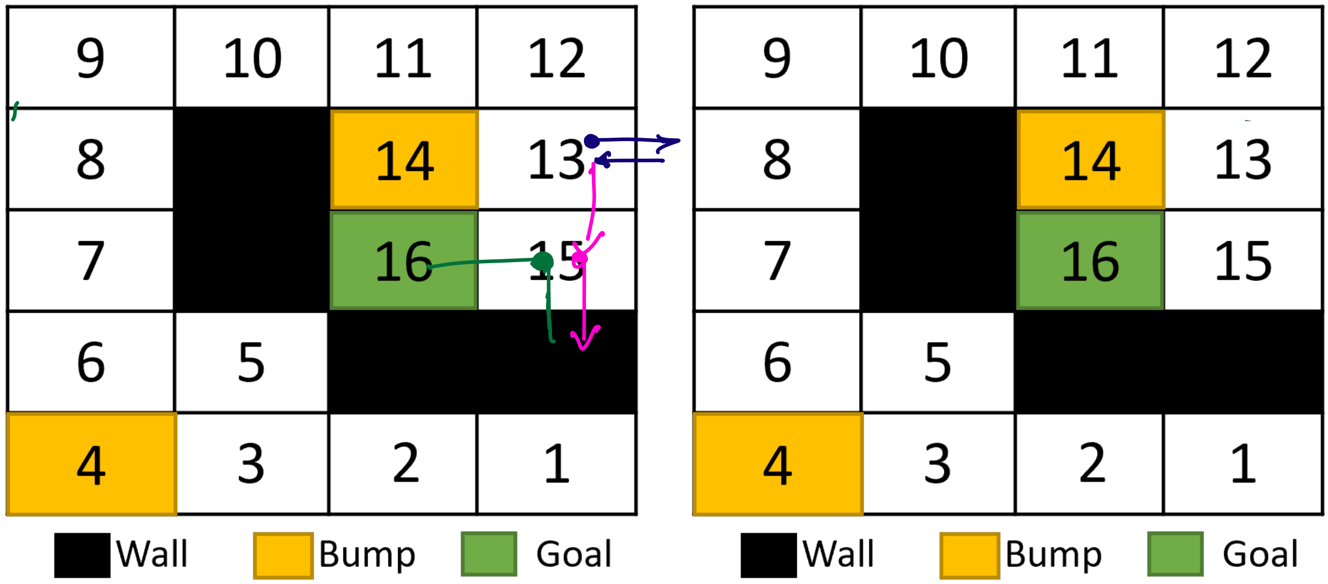
next step $e(s_t,a_t) = \lambda \gamma e(s_t,a_t)$

$0 < \lambda < 1$

↳ Controls the decay rate of eligibility trace







SARSA − λ

episode 1

Random initial state

$(13) \xrightarrow{R = \pi^{\varepsilon g}(13) = \begin{cases} 0.25 \rightarrow U\ D\ R\ L \end{cases}} (13) \xrightarrow{D = \pi^{\varepsilon g}(13) = \begin{cases} 0.25\ U\ D\ R\ L \end{cases}}$

Reward = −1

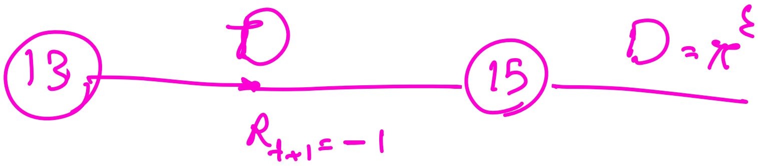
$$e(13, R) = e(13, R) + 1 = 0 + 1 = 1$$

$$\delta_t = R_{t+1} + \gamma\, Q(S_{t+1}, a_{t+1}) - Q(S_t, a_t)$$

$$= -1 + 0.9 \times Q\underbrace{(13, D)}_{0} - \underbrace{Q(13, R)}_{0} = -1$$

$$Q(13, R) = \underbrace{Q(13, R)}_{} + \underbrace{\alpha}_{} \underbrace{\delta_t}_{-1} \underbrace{e(13, R)}_{1} = -0.5$$

$$e(13,R) = \underbrace{e(13,R)}_{1}\,\underbrace{\gamma}_{0.9}\,\underbrace{\lambda}_{0.95} = 0.855$$



$D = \pi^{\epsilon\text{-g}}(13) = \{ 0.25 \ \forall\ D\,R\,L$

$$e(13,D) = e(13,D) + 1 = 1$$

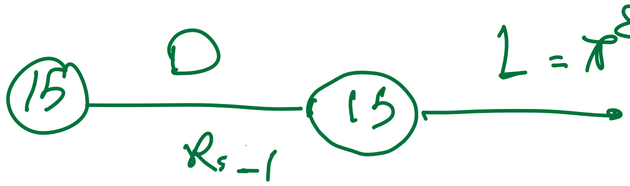
$$\delta_t = -1 + \gamma\,\underbrace{Q(15,L)}_{0} - \underbrace{Q(13,D)}_{0} = -1$$

$$Q(13,D) = \underbrace{Q(13,D)}_{0} + \underbrace{\alpha}_{0.5}\,\underbrace{\delta_t}_{-1}\,\underbrace{e(13,D)}_{1} = -0.5$$

$$Q(13,R) = \underbrace{Q(13,R)}_{-0.5} + \underbrace{\alpha}_{0.5}\,\underbrace{\delta_t}_{-1}\,\underbrace{e(13,R)}_{0.855} = -0.9275$$

$$e(13,D) = \overset{0.9}{\gamma}\,\overset{0.95}{\lambda}\,\overbrace{e(13,D)}^{1} = 0.855$$

$$e(13,R) = \underbrace{\gamma}_{0.9}\,\underbrace{\lambda}_{0.95}\,\underbrace{e(13,R)}_{0.855} = 0.731$$



$L = \pi^{\epsilon g}(15) = \{0.25 \quad U\ R\ L\ D$

$e(15, D) = e(15, D) + 1 = 1$

$\delta_t = -1 + \gamma Q(15, L) - Q(15, D) = -1$

$Q(15, D) = -0.5$

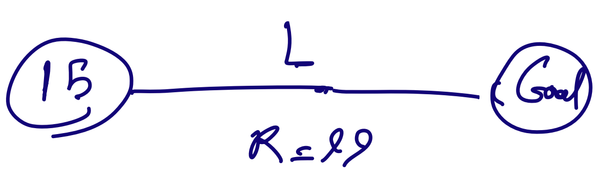
$Q(13, D) = Q(13, D) + \alpha \delta_t \overbrace{e(13, D)}^{0.855} = -0.9275$

$Q(13, R) = Q(13, R) + \alpha \delta_t \underbrace{e(13, R)}_{0.731} = -1.293$

$e(15, D) = \gamma \lambda e(15, D) = 0.855$

$e(13, D) = 0.731$

$e(13, R) = 0.625$



$e(15, L) = e(15, L) + 1 = 1$

$\delta_t = 99 + \gamma \underbrace{Q(\text{Goal}, a)}_{0} - \underbrace{Q(15, L)}_{0} = 99$

$$Q(15,L) = Q(15,L) + \alpha \, \delta_t \, e(15,L) = 49.5$$

$$Q(15,D) = \underbrace{Q(15,D)}_{0} + \underbrace{\alpha}_{0.5} \, \underbrace{\delta_t}_{99} \, \underbrace{e(15,D)}_{0.855} = 41.82$$

$$Q(13,D) = Q(13,D) + \alpha \, \delta_t \, e(13,D) = 35.257$$

$$Q(13,R) = Q(13,R) + \alpha \, \delta_t \, e(13,R) = 29.64$$

SARSA(λ) Algorithm

Initialize $Q(s,a)$ arbitrarily and $e(s,a) = 0$, for all s, a
Repeat (for each episode):
  Initialize s, a $\leftarrow e(s,a) = 0$ for all 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

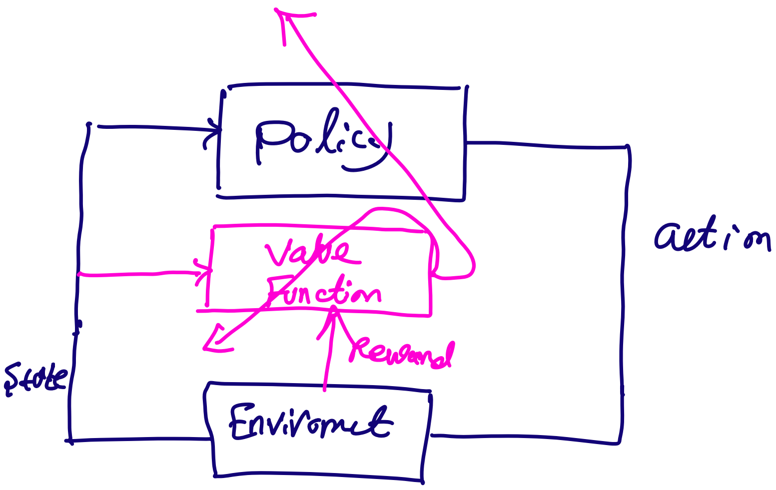
Actor-critic Policy

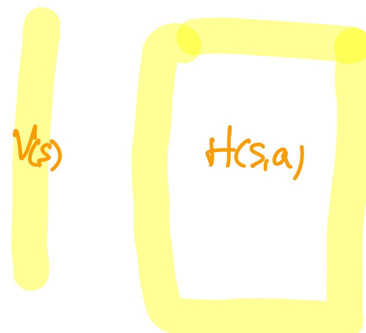
$$\underset{a\in[0,1]}{\arg\max}\ Q(S; a)$$

Q-learning & SARSA $\Longrightarrow$ Value-based policy

$Q \longrightarrow$

$$V(S_t) = V(S_t) + \alpha[\underbrace{R_{t+1} + \gamma V(S_{t+1}) - V(S_t)}_{\delta_t}]$$





$$\delta_t = [R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

$$V(S_t) = V(S_t) + \alpha \delta_t$$

$$\pi(a_t|S_t) = \frac{e^{H(S_t, a_t)}}{\sum_{a\in A} e^{H(S_t, a)}}$$

$$H(S_t, a_t) = H(S_t, a_t) + \beta \delta_t (1 - \pi(a_t|S_t))$$

$\uparrow$ preference

A	B

$$M(a^1) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$M(a^2) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\text{Reward} \begin{cases} B & +5 \\ a^2 & -1 \end{cases}$$

$\alpha = 0.25$, $\beta = 0.3$

$$\boxed{\begin{matrix} H(s,a) = 0 \\ V(s) = 0 \end{matrix}}$$

<u>Episode</u>

Random State

$$S_0 = A \xrightarrow[R=4]{a^2 = \pi(A) =} B \quad \begin{cases} \pi(a^1 \mid A) = \dfrac{e^{H(A,a^1)}}{e^{H(A,a^1)} + e^{H(A,a^2)}} = \dfrac{e^0}{e^0+e^0} = 0.5 & a^1 \\[2ex] \pi(a^2 \mid A) = \dfrac{e^{H(A,a^2)}}{e^{H(A,a^1)} + e^{H(A,a^2)}} = \dfrac{e^0}{e^0+e^0} = 0.5 & a^2 \end{cases}$$

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t) = 4 + \overset{0.9}{\gamma} \cdot \underbrace{V(B)}_{0} - \underbrace{V(A)}_{0} = 4$$

$$\begin{cases} V(\underbrace{S_t}_{A}) = V(\underbrace{S_t}_{A}) + \underbrace{\alpha}_{0.25} \underbrace{\delta_t}_{4} = 1 \\[3ex] H(A,a^2) = \underbrace{H(A,a^2)}_{0} + \underbrace{\beta}_{0.3} \times \underbrace{\delta}_{4} \big(1 - \underbrace{\pi(a^2 \mid A)}_{0.5}\big) = 0.6 \end{cases}$$

$$B \xrightarrow[R=-1]{\pi =} A \quad \begin{cases} \pi(a^1 \mid B) = \dfrac{e^{H(B,a^1)}}{e^{H(B,a^1)} + e^{H(B,a^2)}} = \dfrac{e^0}{e^0+e^0} = 0.5 \\[2ex] \pi(a^2 \mid B) = 1 - \pi(a^1 \mid B) = 0.5 \end{cases} \sim a^2$$

$$\delta_t = -1 + \overset{0.9}{\gamma} \underbrace{V(A)}_{1} - \underbrace{V(B)}_{0} = -0.1$$

$$\begin{cases} V(B) = \underbrace{V(B)}_{0} + \underbrace{\alpha}_{0.25} \underbrace{\delta_t}_{-0.1} = -0.025 \checkmark \\ H(B,a^2) = \underbrace{H(B,a^2)}_{0} + \underbrace{\beta}_{0.3} \underbrace{\delta_t}_{-0.1} \left(1 - \pi(\underbrace{a^2|B}_{0.5})\right) = -0.015 \end{cases}$$

$$A \xrightarrow[R=4]{\pi(a|A)=} B \quad \begin{cases} \pi(a^1|A) = \dfrac{e^{H(A,a^1)}}{e^{H(A,a^1)} + e^{H(A,a^2)}} = \dfrac{e^0}{e^0 + e^{0.6}} = 0.3543 \\ \pi(a^2|A) = 0.6457 \end{cases}$$

$\sim a^2$

$$\delta = \underbrace{R}_{4} + \underbrace{\gamma}_{0.9} \underbrace{V(B)}_{-0.025} - \underbrace{V(A)}_{1} = 2.97$$

$$\begin{cases} V(A) = V(A) + \alpha \delta_t = 1 + 0.25 \times 2.97 = 1.74 \\ H(A,a^2) = H(A,a^2) + \beta \delta \left(1 - \underbrace{\pi(a^2|A)}_{0.6457}\right) = 0.91 \end{cases}$$

$$B \xrightarrow[R=5]{} B \quad \begin{cases} \pi(a^1|B) = 0.5037 \\ \pi(a^2|B) = 0.4963 \end{cases}$$

$\sim a^1$

$$\delta = \underbrace{R}_{5} + \underbrace{\gamma}_{0.9} \underbrace{V(B)}_{-0.025} - V(B) = 5.005$$

$$\begin{cases} V(B) = \underbrace{V(B)}_{-0.025} + \underbrace{\alpha}_{0.25} \underbrace{\delta_t}_{5.005} = 1.225 \\ H(B,a^1) = 0.74 \end{cases}$$

$H(A, a^1) = 0$ $\qquad$ $H(B, a^1) = 0.74$

$H(A, a^2) = 0.916$ $\qquad$ $H(B, a) = -0.015$

$$\pi(a^1 | A) = \frac{e^{H(A, a^1)}}{e^{H(A, a^1)} + e^{H(A, a^1)}} = 0.28$$

$\pi(a^2 | A) = 0.72$

$\pi(a^1 | B) = 0.68$

$\pi(a^2 | B) = 0.31$

Tabular Actor-Critic Algorithm

$V(s)=0$, $H(s,a)=0$, for all $s \in S$, $a \in A$.

Repeat for N episodes

- Start from a random state $s_0 \in S$, $t=0$

  While $t < T$ (episode Length).

  - Select action: $a_t \sim \pi(\cdot \mid s_t)$: $\pi(a|s) = \dfrac{e^{H(s,a)}}{\sum_{a' \in A} e^{H(s,a')}}$
  - Take action a_t, move to state s_{t+1} and observe R_{t+1}.
  - $\delta_t = R_{t+1} + \gamma \, V(s_{t+1}) - V(s_t)$
  - $V(s_t) = V(s_t) + \alpha \, \delta_t$
  - $H(s_t, a_t) = H(s_t, a_t) + \beta \, \delta_t \left(1 - \pi(a_t \mid s_t)\right)$
  - $t = t+1$