Lecture 17 - March 17, 2023

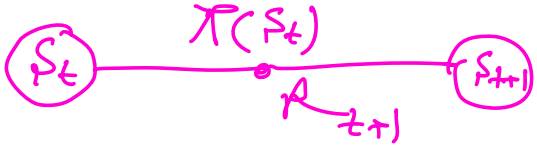
- Temporal Difference Learning
  - TD(0)
  - SARSA
  - Q-Learning
  - On-Policy Vs. Off-Policy
  - Expected SARSA
  - Double Q-Learning
  - Multi-step Bootstrapping
  - SARSA-Lambda
  - Actor-Critic Method

HW3 → Due March 17
Project 3 → Due April 14
HW4 → Due March 31

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

## Review

$$TD \Rightarrow V_\pi(S_t) = V_\pi(S_t) + \alpha\left[\underbrace{R_{t+1} + \gamma V_\pi(S_{t+1})}_{\text{TD Target}} - V_\pi(S_t)\right]$$




$\pi$ Deterministic $\Longrightarrow$ $\pi$ Stochastic $:=$ $\pi^{\varepsilon\text{-greedy}}$

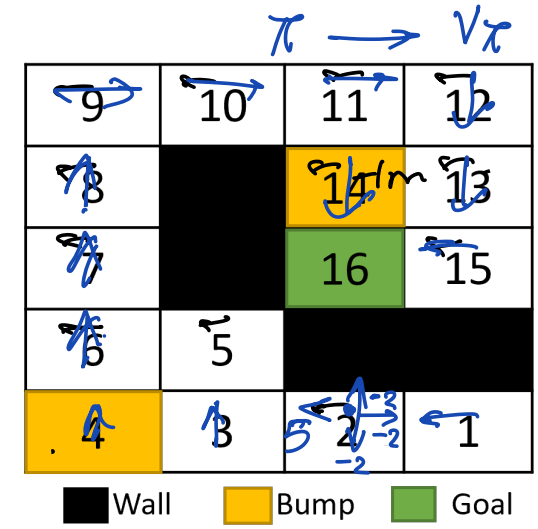
$$Q(S_t, a_t) = Q(S_t, a_t) + \alpha\left[R_{t+1} + \gamma Q(S_{t+1}, a_{t+1}) - Q(S_t, a_t)\right]$$

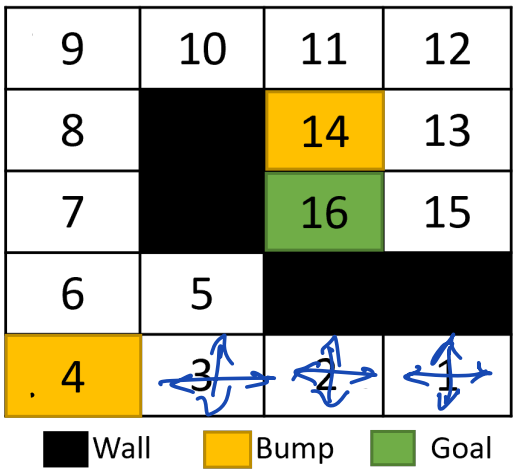
$$\pi^{\varepsilon G}(S) = \begin{cases} \underset{a \in A}{\text{argmax}}\, Q(S, a) & 1-\varepsilon \\ \text{Random}\{a^1, \ldots, a^L\} & \varepsilon \end{cases}$$

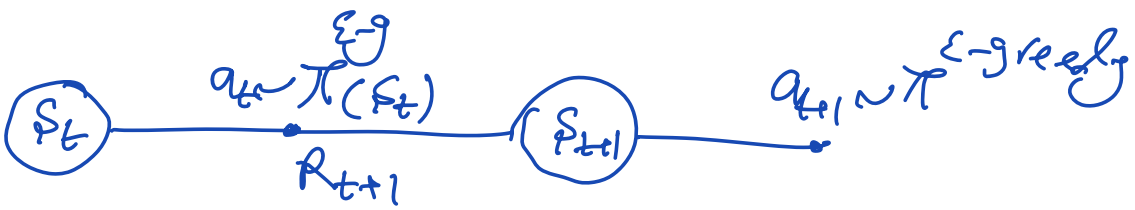
$Q = 0 \xrightarrow{a}$ Update $Q$

$\pi^{\varepsilon G}$ Update

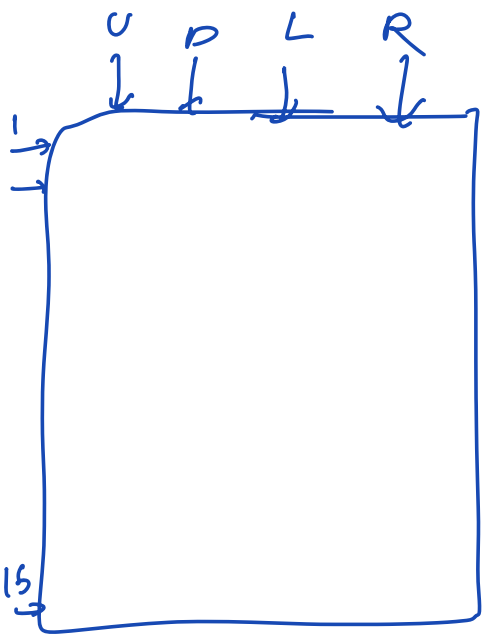
Tabular Learning $\Rightarrow Q$




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

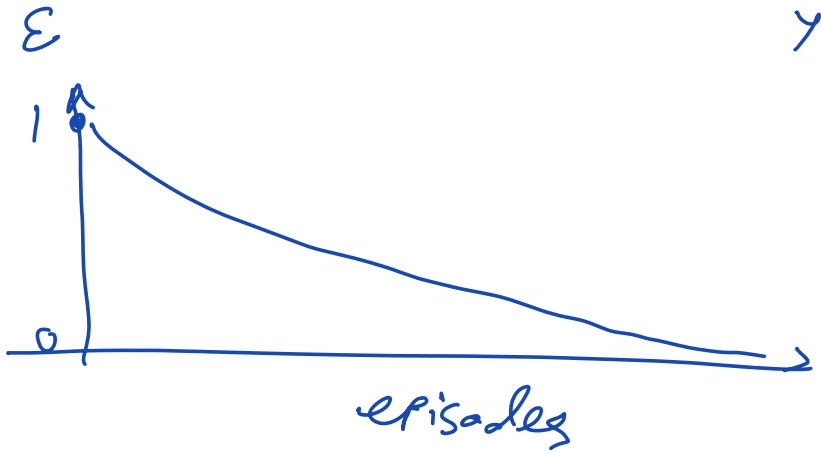



$$\pi^*(s) = \underset{a \in A}{\arg\max}\, Q_{\pi^*}(s, a)$$

The last Q obtained by SARSA is Q for the last $\varepsilon$-g.

$$\pi^*(s) = \underset{a \in A}{\operatorname{argmax}}\, Q^{SARSA}(s, a)$$

You need to follow this when learning has finished and you are in execution




$\varepsilon \xrightarrow{\text{Slowly}} 0$ SARSA $\Rightarrow Q^{SARSA} = Q^{\pi^*}$

# SARSA

$$Q_\pi(s,a) = \mathbb{E}\left[ R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s, a_t = a, a_{t+1:\infty} \sim \pi \right]$$

Bellman $\rightsquigarrow$

$$= \sum_{s'} P(s'|s,a)\left[ R(s,a,s') + \gamma V_\pi(s') \right]$$

$$= \sum_{s'} P(s'|s,a)\left[ R(s,a,s') + \gamma Q_\pi(s', \pi(s')) \right]$$

$P(s'|s,a)$: sample of this

$R(s,a,s') + \gamma Q_\pi(s', \pi(s'))$: SARSA Target

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

$Q(s_t,a_t)$ (left): new

$Q(s_t,a_t)$ (right): old

$R_{t+1} + \gamma Q(s_{t+1}, a_{t+1})$: SARSA Target

$Q(s_t,a_t)$ (last): old

---

# Q-Learning

$$Q_{\pi^*}(s,a) = \mathbb{E}\left[ R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s, a_t = a, a_{t+1:\infty} \sim \pi^* \right]$$

$$= \sum_{s'} P(s'|s,a)\left[ R(s,a,s') + \gamma V_{\pi^*}(s') \right]$$

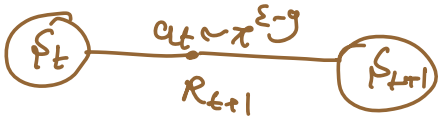
$$= \sum_{s'} P(s'|s,a)\left[ R(s,a,s') + \gamma Q_{\pi^*}(s', \pi^*(s')) \right]$$

$$= \sum_{s'} P(s'|s,a)\left[ R(s,a,s') + \gamma \max_{\hat{a} \in A} Q_{\pi^*}(s', \hat{a}) \right]$$

$R(s,a,s') + \gamma \max_{\hat{a} \in A} Q_{\pi^*}(s', \hat{a})$: Q-Learning Target

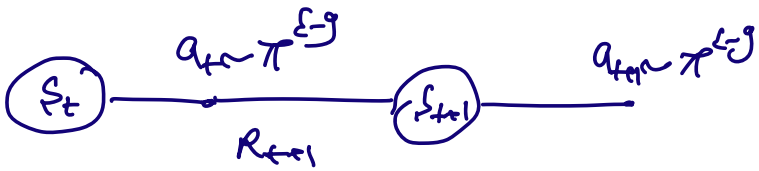
## Q-Learning

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




## SARSA

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



## Q-Learning

Initialize $Q(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
    Initialize $S$
    Repeat (for each step of episode):
        Choose $A$ from $S$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
        Take action $A$, observe $R$, $S'$
        $Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma \max_a Q(S',a) - Q(S,A)]$
        $S \leftarrow S'$;
    until $S$ is terminal

| A | B |
|---|---|

B → 5
$a^2$ → -1

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

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

Q-Learning ↓

$\varepsilon$-greedy $\pi(A) = \begin{cases} \text{argmax}_a Q(A,a) & 1-\varepsilon \\ \text{Random} & \varepsilon \end{cases}$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

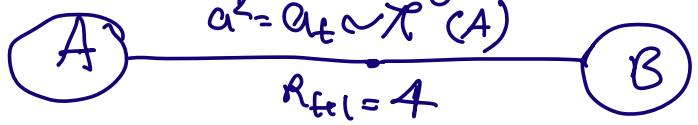
Closet Deterministic Policy to $\pi^{\varepsilon\text{-greedy}}$

$\pi(A) = a^1$ or $a^2$ $\quad$ $\pi(B) = a^1$ or $a^2$

Episode 1

Random State

$a^2 = a_t \sim \pi^{\varepsilon\text{-greedy}}(A)$ → $\begin{cases} 0.5 & a^1 \\ 0.5 & a^2 \end{cases}$

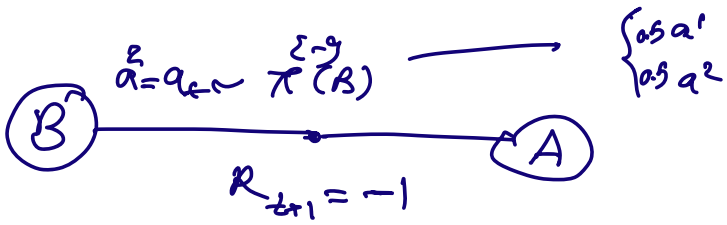



$$Q(A,a^2) = \underbrace{Q(A,a^2)}_{0} + \underbrace{\alpha}_{0.5}\left[\underbrace{R_{t+1}}_{4} + \gamma \max_{a \in A} \underbrace{Q(B,a)}_{0} - \underbrace{Q(A,a^2)}_{0}\right] = 2$$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

Closet Deterministic Policy to $\pi^{\varepsilon\text{-greedy}}$

$\pi(A) = a^2$ $\quad$ $\pi(B) = a^1$ or $a^2$



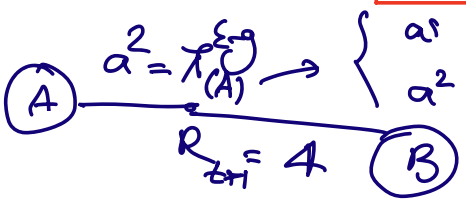

$$Q(B,a^2) = \underbrace{Q(B,a^2)}_{0} + \alpha\Big[\underbrace{R_{t+1}}_{-1} + \gamma \underbrace{\max_{a\in A} Q(A,a)}_{2} - \underbrace{Q(B,a^2)}_{0}\Big]$$

$$= 0.4$$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 2 \\ 0.4 \end{bmatrix}$$

Closest Deterministic Policy to $\pi^{\varepsilon\text{-}g}$

$\pi(A) =$ $a^2$ $\quad$ $\pi(B) =$ $a^2$



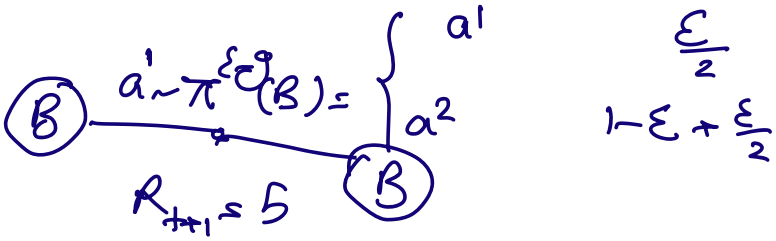

$$Q(A,a^2) = \underbrace{Q(A,a^2)}_{2} + \alpha\Big[\underbrace{R_{t+1}}_{4} + \gamma \underbrace{\max_{a\in A} Q(B,a)}_{0.4} - \underbrace{Q(A,a^2)}_{2}\Big]$$

$$= 3.18$$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 3.18 \\ 0.4 \end{bmatrix}$$

Closet Deterministic Policy to $\pi^{\epsilon-g}$

$\pi(A) =$ $a^2$ $\pi(B) =$ $a^2$




$$Q(B,a^1) = \underbrace{Q(B,a^1)}_{0} + \underbrace{\alpha}_{0.5}\Big[\underbrace{R_{t+1}}_{5} + \gamma \max_{a\in A} \underbrace{Q(B,a)}_{0.4} - \underbrace{Q(B,a^1)}_{0}\Big]$$

$= 2.68$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 2.68 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 3.18 \\ 0.4 \end{bmatrix}$$

Closet Deterministic Policy to $\pi^{\epsilon-g}$

$\pi(A) =$ $a^2$ $\pi(B) = a^1$

Greedy

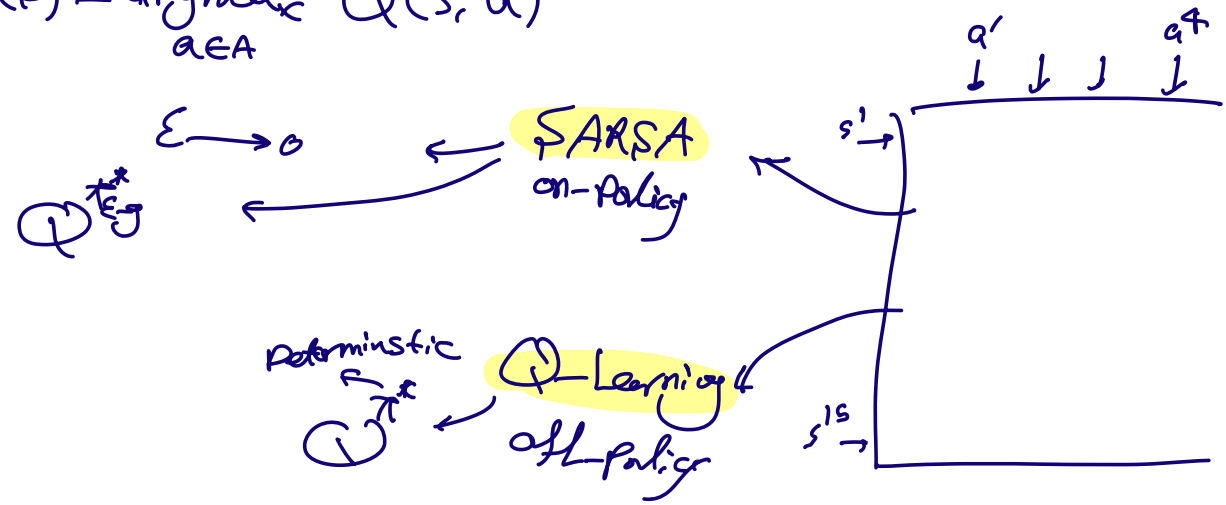
$$\pi(s) = \underset{a\in A}{\arg\max}\, Q(s,a)$$

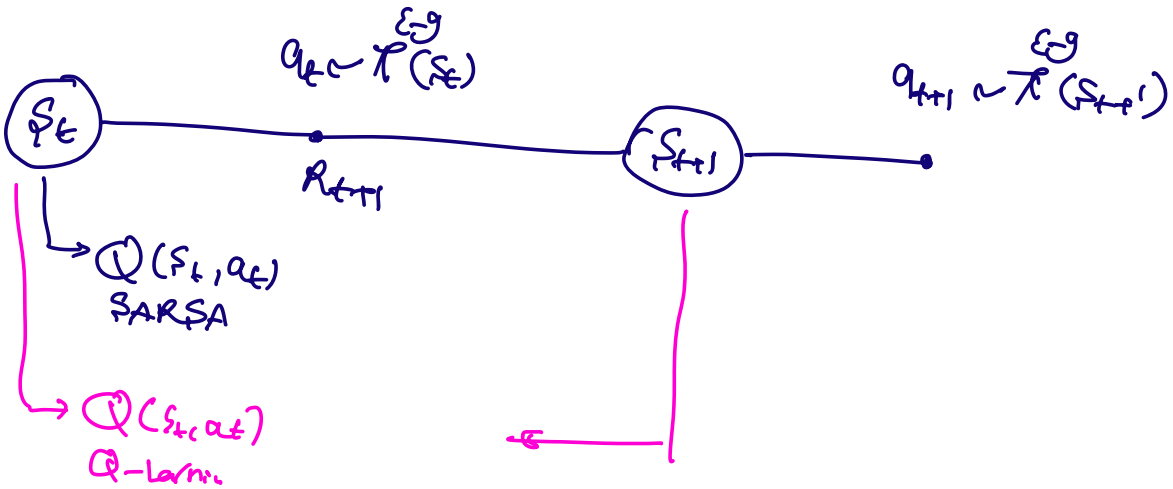
# off-policy Vs on-policy

$$a_t \sim \pi^{\varepsilon\text{-g}}: \begin{cases} \text{greedy} & 1-\varepsilon \\ \text{Random} & \varepsilon \end{cases} = \begin{cases} \underset{a}{\text{argmax}}\, Q(s_t, a) & 1-\varepsilon \\ \text{Random} & \varepsilon \end{cases}$$

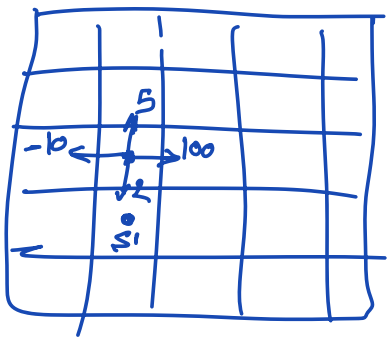
last

$$\pi(s) = \underset{a \in A}{\text{argmax}}\, Q(s, a)$$



$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha\left[R_{t+1} + \gamma \underbrace{\max_{a'} Q(s_{t+1}, a')}_{\text{SARSA: } Q(s_{t+1}, a_{t+1})} - Q(s_t, a_t)\right]$$




$$\pi^{\varepsilon\text{-greedy}}(s_{t+1}) = \begin{cases} R & 1-\varepsilon+\frac{\varepsilon}{4} \\ L & \frac{\varepsilon}{4} \\ D & \frac{\varepsilon}{4} \\ U & \frac{\varepsilon}{4} \end{cases}$$

Q-learning $\max\{100, 5, -10, 2\}$

$$Q(s', U) = Q(s, U) + \alpha\left[\overbrace{R}^{-1} + \gamma \underbrace{\quad\quad}_{} - Q(s, U)\right]$$

$$(1-\varepsilon+\tfrac{\varepsilon}{4})(100) + \tfrac{\varepsilon}{4}(-10) + \tfrac{\varepsilon}{4}(5) + \tfrac{\varepsilon}{4}(2)$$

| | |
|---|---|
| 100 | $1-\varepsilon+\frac{\varepsilon}{4}$ |
| -10 | $\frac{\varepsilon}{4}$ |
| 5 | $\frac{\varepsilon}{4}$ |
| 2 | $\frac{\varepsilon}{4}$ |

$$\pi^{SARSA}(s) = \arg\max_a Q^{SARSA}(s, a) \neq \pi^{Q\text{-learning}}(s) = \arg\max_a Q^{Q\text{-learning}}(s, a)$$

If you are using off-policy $\Rightarrow$ you can explore as you want

Batch learning

for on-policy $\longrightarrow$ $\varepsilon$ should be small or approach to small value

## SARSA

Initialize $Q(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
  Initialize $S$
  Choose $A$ from $S$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma Q(S',A') - Q(S,A)]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

## Q-Learning

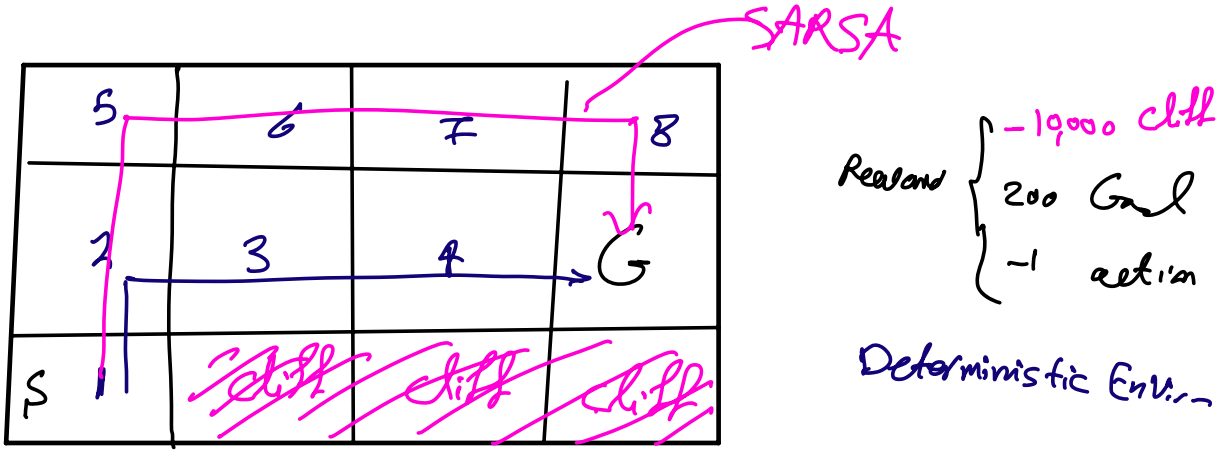
Initialize $Q(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
  Initialize $S$
  Repeat (for each step of episode):
    Choose $A$ from $S$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma \max_a Q(S',a) - Q(S,A)]$
    $S \leftarrow S'$;
  until $S$ is terminal

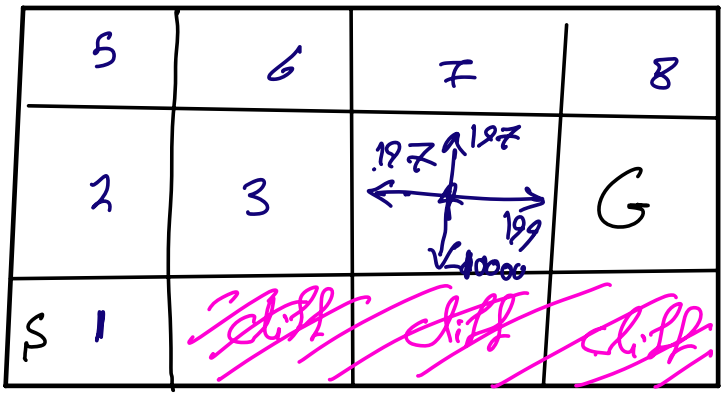
SARSA

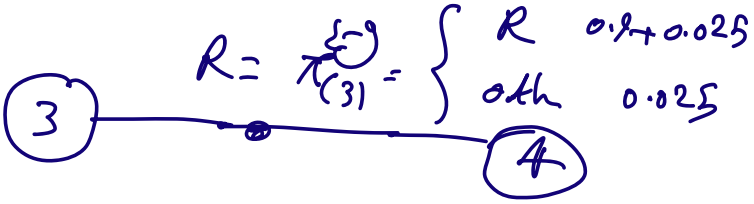
$$Q(S_t,a_t) = Q(S_t,a_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1},a_{t+1}) - Q(S_t,a_t)]$$

Q Learning

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

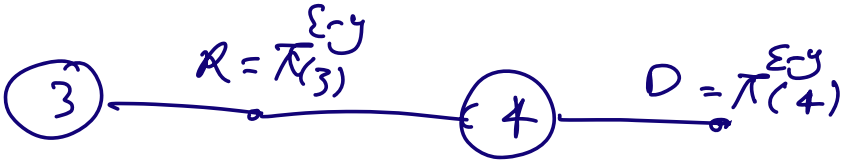



$$Q(3,R) = Q(3,R) + \alpha\left[\underbrace{R}_{-1} + \gamma \max_{a \in A} Q(4,a) - Q(3,R)\right]$$

Q-learning



$\pi(4) = \begin{cases} R & 0.925 \rightarrow 199 \\ L & 0.025 \quad 197 \\ U & 0.025 \quad 197 \\ D & 0.025 \quad -10000 \end{cases}$

SARSA

$Q(3,R) = Q(3,R) + \alpha[\underbrace{R}_{-1} + \gamma Q(4,D) - Q(3,R)]$