Deep Reinforcement Learning

2022-23 Second Semester, M.Tech (AIML)

# Session #9:
# Temporal Difference Learning

S.P.Vimal | Dept. of CSIS, WILP, BITS-Pilani | vimalsp@wilp.bits-pilani.ac.in

# Agenda for the class

- Temporal Difference Learning
  - TD(0)
  - SARSA
  - Q-Learning

Acknowledgements: Some of the slides were adopted *with permission* from the course CSCE-689 (Texas A&M University) by Prof. Guni Sharon

# Solving MDPs so far

**Dynamic programming**

- ☑ Off policy
- ☑ local learning, propagating values from neighbors (Bootstrapping)
- ✗ Model based

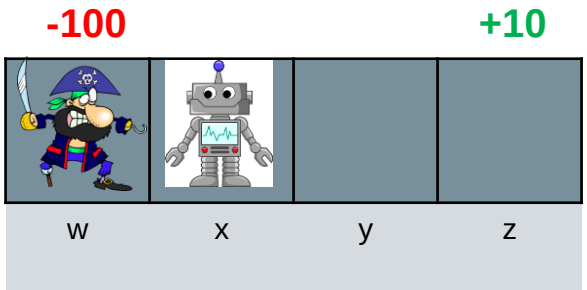
**Monte-Carlo**

- ✗ On-policy (though important sampling can be used)
- ✗ Requires a full episode to train on
- ☑ Model free, online learning

- $Q(z, exit) = 10$
- $Q(y, \rightarrow) = 0 + \gamma \max_a Q(z, a)$
- $Q(x, \rightarrow) = 0 + \gamma \max_a Q(y, a)$

$$q^*(s,a) = \sum_{s'} p(s'|s,a)\,(r(s,a,s') + \gamma \max_a [q^*(s',a)])$$

$\gamma = 0.9$




- Episode = $\{x, y, z, exit\}$
- $Q(z, exit) = 10$
- $Q(y, \rightarrow) = 9$
- $Q(x, \rightarrow) = 8.1$

# Fuse DP and MC

**Dynamic programming**

- ☑ Off policy
- ☑ local learning, propagating values from neighbors (Bootstrapping)
- ✗ Model based

**Monte-Carlo**

- ✗ On-policy (though important sampling can be used)
- ✗ Requires a full episode to train on
- ☑ Model free, online learning

**TD Learning**

- ☑ Off policy
- ☑ local learning, propagating values from neighbors (Bootsraping)
- ☑ Model free, online learning

# Temporal difference learning

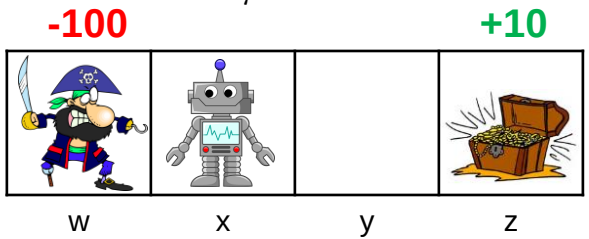
- 
- $Q(s,a) = Q(s,a) + \alpha\left(R_{t+1} + \gamma \max_{a'}[Q(s',a')] - Q(s,a)\right)$
  - $V(s) = V(s) + \alpha\big(R_{t+1} + \gamma V(s') - V(s)\big)$
- Update estimate based on other estimates
- Model free
- Online, incremental learning
- Guaranteed to converge to the true value!
  - Some conditions on the step size, $\alpha$ (see slide #19 in 2Multi_armed_bandits.pptx)
- Usually converges faster than MC methods

# SARSA: On-policy TD Control

**Sarsa (on-policy TD control) for estimating $Q \approx q_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

$\gamma = 0.9$

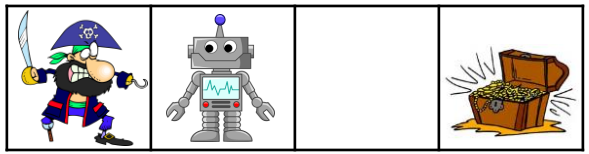


# SARSA: On-policy TD Control

**Sarsa (on-policy TD control) for estimating $Q \approx q_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

$\gamma = 0.9$




$$Q = \begin{array}{|c|c|c|c|} \hline 0 & 0{,}0 & 0{,}0 & 0 \\ \hline \end{array}$$

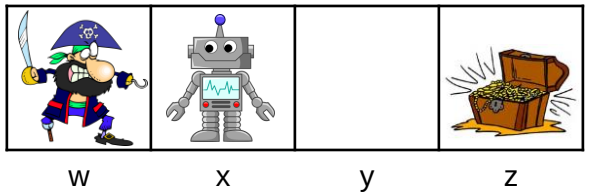
w x y z

# SARSA: On-policy TD Control

**Sarsa (on-policy TD control) for estimating $Q \approx q_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

$\gamma = 0.9$




| $x$ | ← | 0 | $w$ | $null$ |
|---|---|---|---|---|
| S | A | R | S' | A' |

$$Q = \begin{array}{|c|c|c|c|}\hline 0 & 0{,}0 & 0{,}0 & 0 \\ \hline \end{array}$$

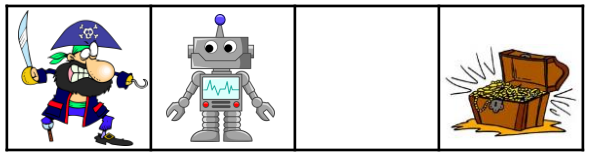
w x y z

# SARSA: On-policy TD Control

**Sarsa (on-policy TD control) for estimating $Q \approx q_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

$\gamma = 0.9$




| x | ← | 0 | w | exit |
|---|---|---|---|---|
| S | A | R | S' | A' |

$$Q = \begin{array}{|c|c|c|c|} \hline 0 & 0{,}0 & 0{,}0 & 0 \\ \hline \end{array}$$

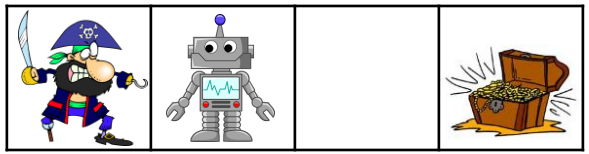
w x y z

# SARSA: On-policy TD Control

**Sarsa (on-policy TD control) for estimating $Q \approx q_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

$\gamma = 0.9$




| $x$ | ← | 0 | $w$ | $exit$ |
|---|---|---|---|---|
| S | A | R | S' | A' |

$$Q = \begin{array}{|c|c|c|c|}\hline 0 & 0{,}0 & 0{,}0 & 0 \\ \hline \end{array}$$

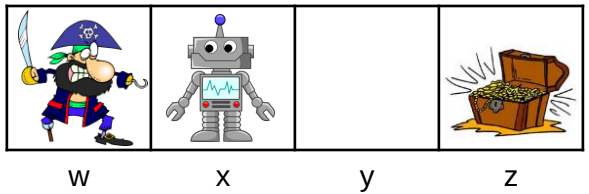
w x y z

# SARSA: On-policy TD Control

**Sarsa (on-policy TD control) for estimating $Q \approx q_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

$\gamma = 0.9$




| *w* | *exit* | 0 | *w* | *exit* |
|---|---|---|---|---|
| S | A | R | S’ | A’ |

$$Q = \begin{array}{|c|c|c|c|} \hline \textbf{0} & \textbf{0,0} & \textbf{0,0} & \textbf{0} \\ \hline \end{array}$$

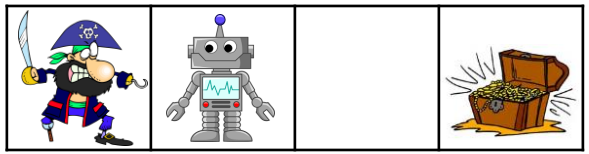
w x y z

# SARSA: On-policy TD Control

**Sarsa (on-policy TD control) for estimating $Q \approx q_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma Q(S',A') - Q(S,A)\right]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

$\gamma = 0.9$




| *w* | *exit* | −100 | *ter* | *exit* |
|---|---|---|---|---|
| S | A | R | S’ | A’ |

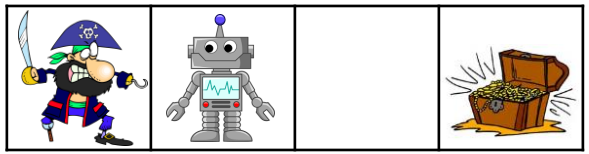
$$Q = \begin{array}{|c|c|c|c|} \hline \textbf{0} & \textbf{0,0} & \textbf{0,0} & \textbf{0} \\ \hline \textbf{w} & \textbf{x} & \textbf{y} & \textbf{z} \\ \hline \end{array}$$

# SARSA: On-policy TD Control

**Sarsa (on-policy TD control) for estimating $Q \approx q_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

$\gamma = 0.9$




| $w$ | $exit$ | $-100$ | $ter$ | · |
|---|---|---|---|---|
| S | A | R | S' | A' |

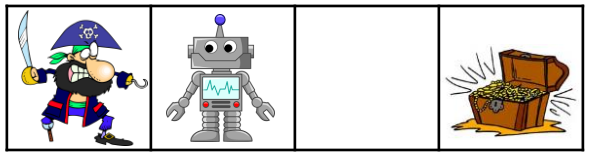
$$Q = \begin{array}{|c|c|c|c|} \hline \textbf{-100} & \textbf{0,0} & \textbf{0,0} & \textbf{0} \\ \hline \textbf{w} & \textbf{x} & \textbf{y} & \textbf{z} \\ \hline \end{array}$$

# SARSA: On-policy TD Control

**Sarsa (on-policy TD control) for estimating $Q \approx q_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

$\gamma = 0.9$




| $x$ | ← | $-100$ | $ter$ | · |
|---|---|---|---|---|
| S | A | R | S’ | A’ |

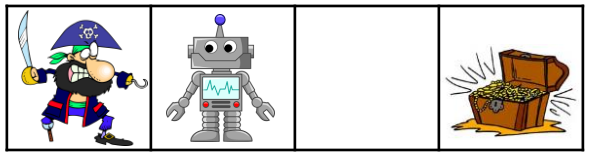
$$Q = \begin{array}{|c|c|c|c|} \hline \textbf{-100} & \textbf{0,0} & \textbf{0,0} & \textbf{0} \\ \hline \textbf{w} & \textbf{x} & \textbf{y} & \textbf{z} \\ \hline \end{array}$$

# SARSA: On-policy TD Control

**Sarsa (on-policy TD control) for estimating $Q \approx q_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

$\gamma = 0.9$




| $x$ | ← | 0 | $w$ | $exit$ |
|---|---|---|---|---|
| S | A | R | S' | A' |

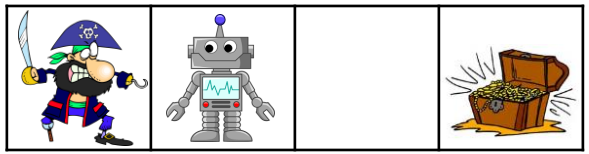
$$Q = \begin{array}{|c|c|c|c|}\hline \textbf{-100} & \textbf{0,0} & \textbf{0,0} & \textbf{0} \\ \hline \textbf{w} & \textbf{x} & \textbf{y} & \textbf{z} \\ \hline \end{array}$$

# SARSA: On-policy TD Control

**Sarsa (on-policy TD control) for estimating $Q \approx q_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

$\gamma = 0.9$




| x | ← | 0 | w | exit |
|---|---|---|---|---|
| S | A | R | S' | A' |

$$Q = \begin{array}{|c|c|c|c|} \hline \text{-100} & \text{-90,0} & \text{0,0} & \text{0} \\ \hline \text{w} & \text{x} & \text{y} & \text{z} \\ \hline \end{array}$$

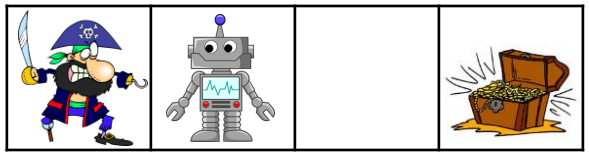
# SARSA: On-policy TD Control

**Sarsa (on-policy TD control) for estimating $Q \approx q_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

And so on…

$\gamma = 0.9$




| *w* | *exit* | 0 | *w* | *exit* |
|---|---|---|---|---|
| S | A | R | S’ | A’ |

$$Q = \begin{array}{|c|c|c|c|} \hline \textbf{-100} & \textbf{-90,0} & \textbf{0,0} & \textbf{0} \\ \hline \textbf{w} & \textbf{x} & \textbf{y} & \textbf{z} \\ \hline \end{array}$$

# Q-learning: Off-policy TD Control

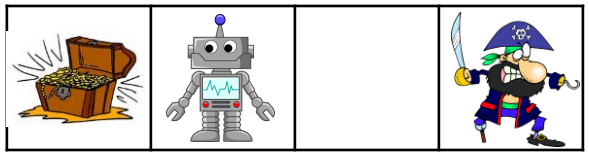
- Use the original TD update rule
- $Q(s,a) = Q(s,a) + \alpha\left(R_{t+1} + \gamma \max_{a'}[Q(s',a')] - Q(s,a)\right)$
- Approximates the state-action value for the optimal policy, i.e., $q^*$
  - Assuming that every state-action pair is visited infinitely often
- Follows from the proof of convergence for the Bellman function
  - See slides MDPs+DP

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| S | A | R | S' |
|---|---|---|---|
| *x* | *null* | *null* | *null* |

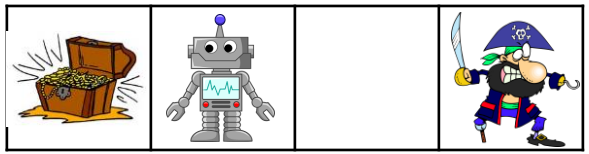
$$Q = \begin{array}{|c|c|c|c|} \hline \textbf{0} & \textbf{0,0} & \textbf{0,0} & \textbf{0} \\ \hline \textbf{w} & \textbf{x} & \textbf{y} & \textbf{z} \\ \hline \end{array}$$

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| $x$ | → | 0 | $y$ |
|---|---|---|---|
| S | A | R | S’ |

$$Q = \begin{array}{|c|c|c|c|}\hline 0 & 0{,}0 & 0{,}0 & 0 \\ \hline \end{array}$$

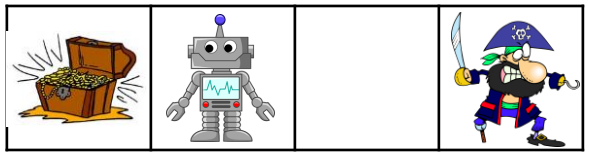
w x y z

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| $x$ | $\rightarrow$ | **0** | $y$ |
|---|---|---|---|
| S | A | R | S’ |

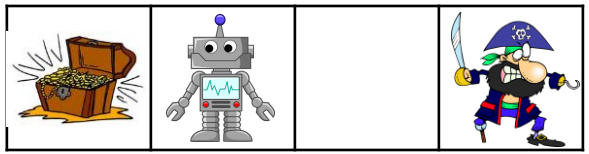
$$Q = \begin{array}{|c|c|c|c|} \hline \textbf{0} & \textbf{0,0} & \textbf{0,0} & \textbf{0} \\ \hline \text{w} & \text{x} & \text{y} & \text{z} \\ \hline \end{array}$$

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| y | → | 0 | y |
|---|---|---|---|
| S | A | R | S' |

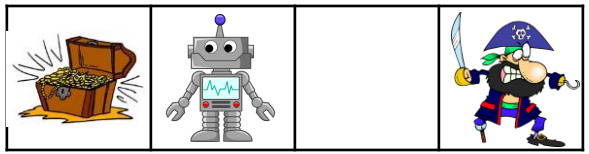
$$Q = \begin{array}{|c|c|c|c|} \hline 0 & 0{,}0 & 0{,}0 & 0 \\ \hline w & x & y & z \\ \end{array}$$

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| y | → | 0 | z |
|---|---|---|---|
| S | A | R | S’ |

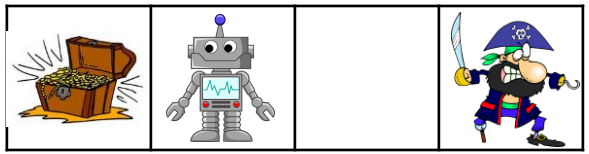
$$Q = \begin{array}{|c|c|c|c|} \hline 0 & 0{,}0 & 0{,}0 & 0 \\ \hline \text{w} & \text{x} & \text{y} & \text{z} \\ \end{array}$$

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| z | → | 0 | z |
|---|---|---|---|
| S | A | R | S’ |

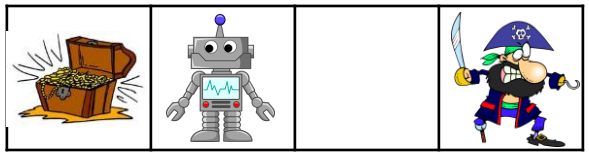
$$Q = \begin{array}{|c|c|c|c|} \hline 0 & 0{,}0 & 0{,}0 & 0 \\ \hline \text{w} & \text{x} & \text{y} & \text{z} \\ \hline \end{array}$$

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| $z$ | $exit$ | $-100$ | $\cdot$ |
|---|---|---|---|
| S | A | R | S’ |

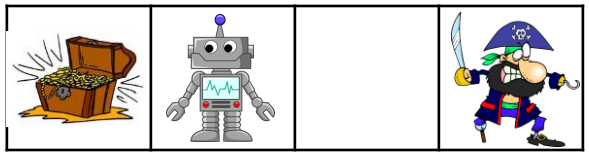
$$Q = \begin{array}{|c|c|c|c|} \hline \textbf{0} & \textbf{0,0} & \textbf{0,0} & \textbf{0} \\ \hline \textbf{w} & \textbf{x} & \textbf{y} & \textbf{z} \\ \hline \end{array}$$

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| $z$ | $exit$ | $-100$ | $\cdot$ |
|---|---|---|---|
| S | A | R | S’ |

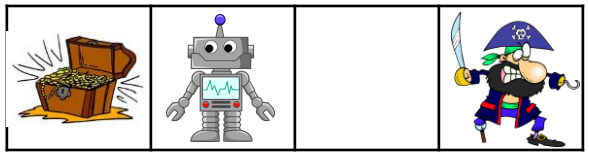
$$Q = \begin{array}{|c|c|c|c|} \hline \textbf{0} & \textbf{0,0} & \textbf{0,0} & \textbf{-100} \\ \hline \textbf{w} & \textbf{x} & \textbf{y} & \textbf{z} \\ \hline \end{array}$$

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| $x$ | $exit$ | $-100$ | $\cdot$ |
|---|---|---|---|
| S | A | R | S’ |

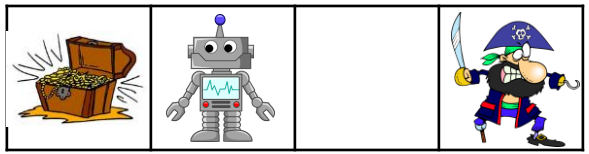
$$Q = \begin{array}{|c|c|c|c|} \hline \textbf{0} & \textbf{0,0} & \textbf{0,0} & \textbf{-100} \\ \hline \textbf{w} & \textbf{x} & \textbf{y} & \textbf{z} \\ \hline \end{array}$$

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| $x$ | $\rightarrow$ | **0** | $y$ |
|---|---|---|---|
| S | A | R | S’ |

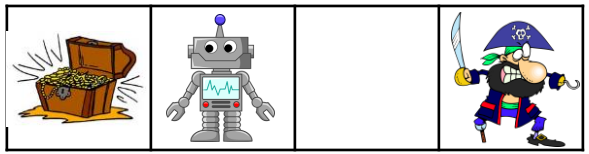
$$Q = \begin{array}{|c|c|c|c|} \hline \textbf{0} & \textbf{0,0} & \textbf{0,0} & \textbf{-100} \\ \hline \textbf{w} & \textbf{x} & \textbf{y} & \textbf{z} \\ \hline \end{array}$$

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| y | → | 0 | z |
|---|---|---|---|
| S | A | R | S’ |

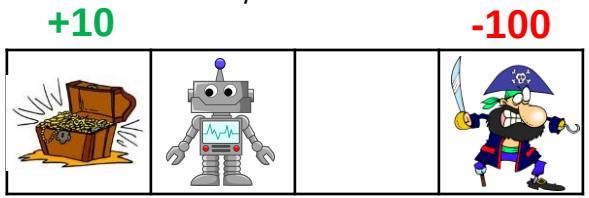
$$Q = \begin{array}{|c|c|c|c|} \hline 0 & 0{,}0 & 0{,}0 & -100 \\ \hline \text{w} & \text{x} & \text{y} & \text{z} \\ \end{array}$$

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| y | → | 0 | z |
|---|---|---|---|
| S | A | R | S' |

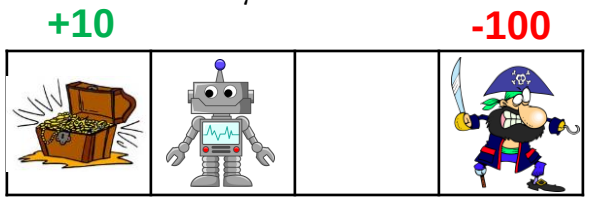
$$Q = \begin{array}{|c|c|c|c|} \hline 0 & 0{,}0 & 0{,}-90 & -100 \\ \hline \text{w} & \text{x} & \text{y} & \text{z} \\ \hline \end{array}$$

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| $x$ | → | 0 | $z$ |
|---|---|---|---|
| S | A | R | S’ |

$$Q = \begin{array}{|c|c|c|c|} \hline 0 & 0{,}0 & 0{,}-90 & -100 \\ \hline \end{array}$$

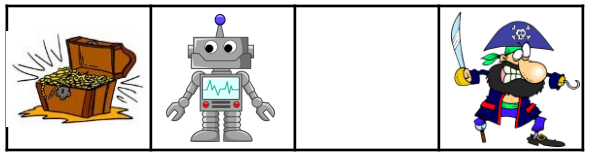
w x y z

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| $x$ | → | 0 | $y$ |
|---|---|---|---|
| S | A | R | S’ |

$Q =$

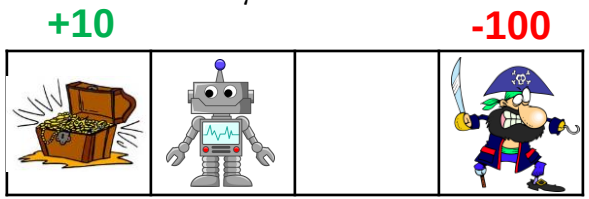
| 0 | 0,0 | 0,-90 | -100 |
|---|---|---|---|
| w | x | y | z |

# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

$\gamma = 0.9$




| $x$ | $\rightarrow$ | **0** | $y$ |
|---|---|---|---|
| S | A | R | S’ |

$$Q = \begin{array}{|c|c|c|c|} \hline \textbf{0} & \textbf{0,0} & \textbf{0,-90} & \textbf{-100} \\ \hline \textbf{w} & \textbf{x} & \textbf{y} & \textbf{z} \\ \hline \end{array}$$

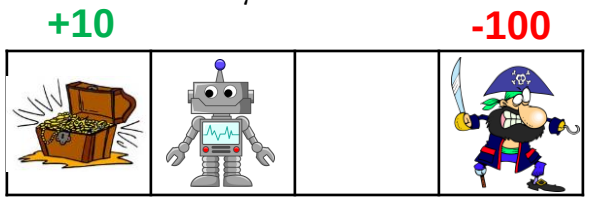
# Q-learning: Off-policy TD Control

**Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$**

Initialize $Q(s,a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, 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., $\epsilon$-greedy)
    Take action $A$, observe $R$, $S'$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$
  until $S$ is terminal

And so on…

$\gamma = 0.9$




| $x$ | → | 0 | $y$ |
|---|---|---|---|
| S | A | R | S’ |

$$Q = \begin{array}{|c|c|c|c|} \hline 0 & 0,0 & 0,-90 & -100 \\ \hline w & x & y & z \\ \end{array}$$

# Required Readings

1. Chapter-6 of Introduction to Reinforcement Learning,2nd Ed., Sutton & Barto

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Thank you