BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Deep Reinforcement Learning

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

Session #6-7:
Monte Carlo Methods

Instructors :

1. Prof. S. P. Vimal (vimalsp@wilp.bits-pilani.ac.in),
2. Prof. Sangeetha Viswanathan (sangeetha.viswanathan@pilani.bits-pilani.ac.in)

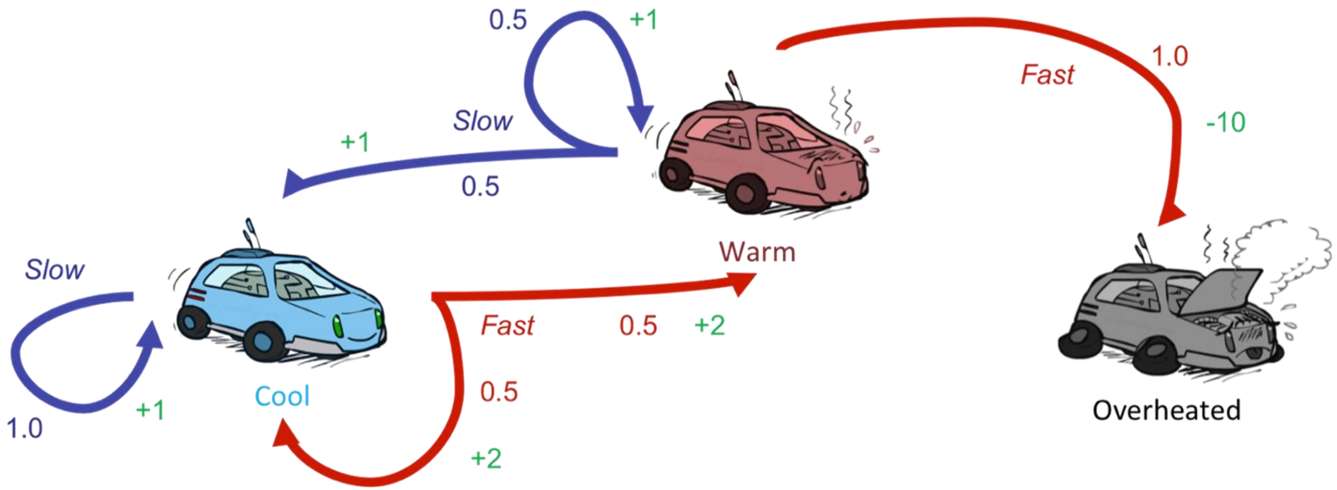
Agenda for the class

- Introduction
- On-Policy Monte Carlo Methods
- Off-Policy Monte Carlo Methods

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

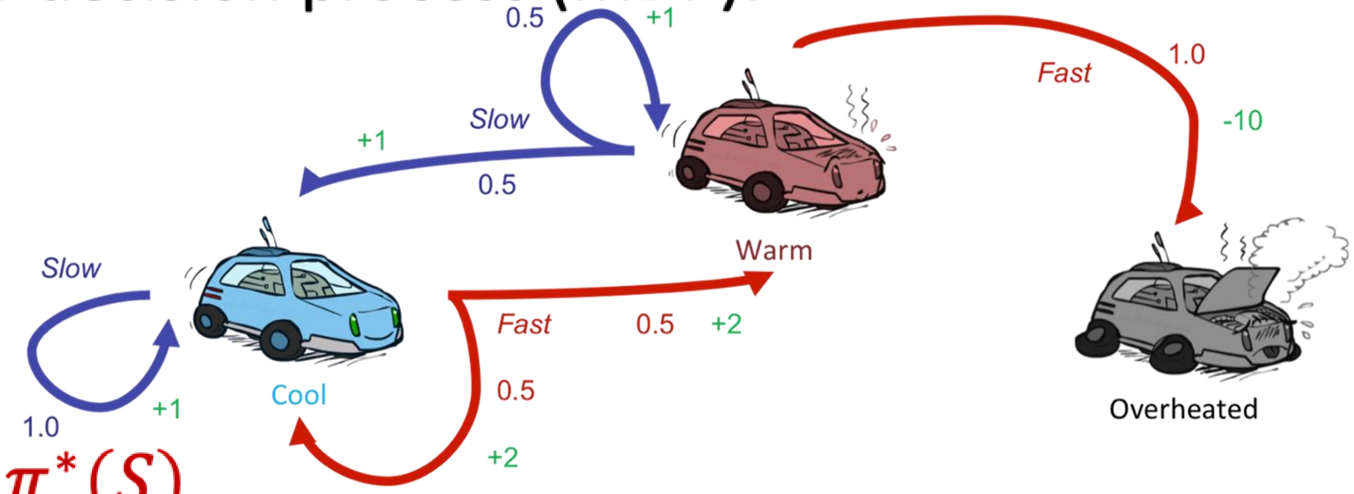
Introduction

- Recollect the problem
  - We need to learn a policy that takes us as far and as faster possible;



Introduction

- Still assume an underlying Markov decision process (MDP):
  - A set of states s ∈ S
  - A set of actions A
  - A model $P(s'|s,a)$
  - A reward function $R(s,a,s')$
  - A discount factor γ
  - Still looking for the best policy $\pi^*(S)$

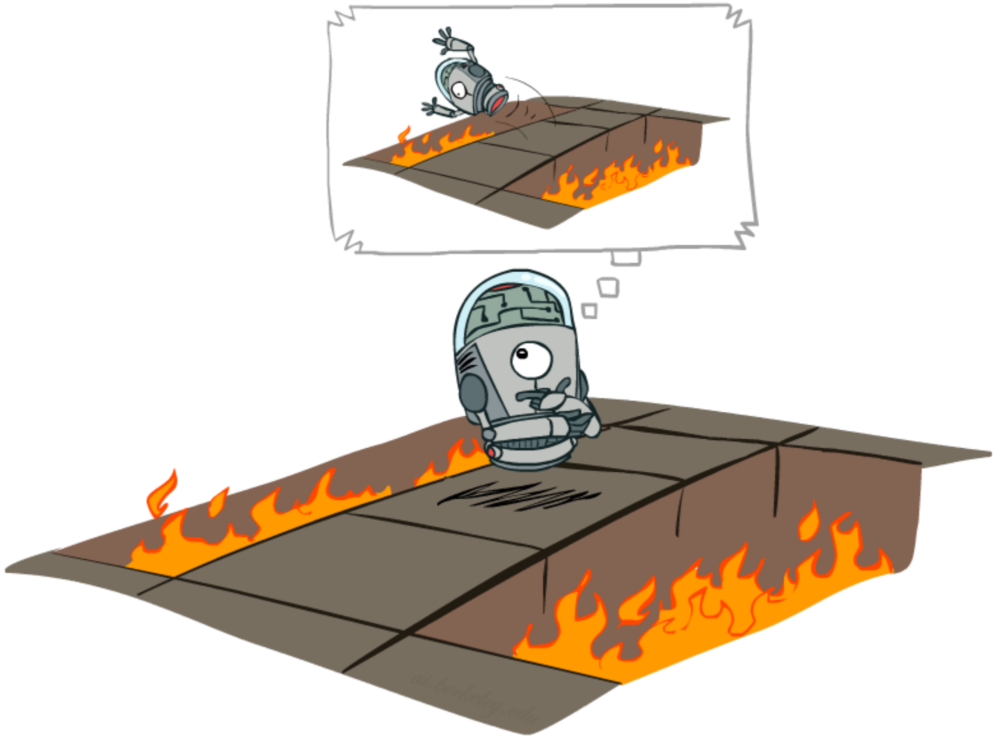


Introduction

- Still assume an underlying Markov decision process (MDP):
  - A set of states s ∈ S
  - A set of actions A
  - A model $P(s'|s,a)$
  - A reward function $R(s,a,s')$
  - A discount factor γ
  - Still looking for the best policy $\pi^*(S)$

- New twist: don't know the model and the reward function
  - That is, we don't know the actions' outcome
  - Must interact with the environment to learn

(Aside) Offline vs. Online (RL)



Offline Optimization

Online Learning

Monte Carlo Methods

- Monte Carlo methods are a broad class of computational algorithms that *rely on repeated random sampling to obtain numerical results*
- The underlying concept is to obtain unbiased samples from a complex/unknown distribution through a random process
- They are often used in physical and mathematical problems and are most useful when it is difficult or impossible to compute a solution analytically
  - Weather prediction
  - Computational biology
  - Computer graphics
  - Finance and business
  - Sport game prediction

First-visit Monte-Carlo Policy Evaluation [estimate Vπ(s)]

Initialize:

$\pi \leftarrow$ policy to be evaluated

$V \leftarrow$ an arbitrary state-value function

$Returns(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$

Repeat forever:

(a) Generate an episode using π

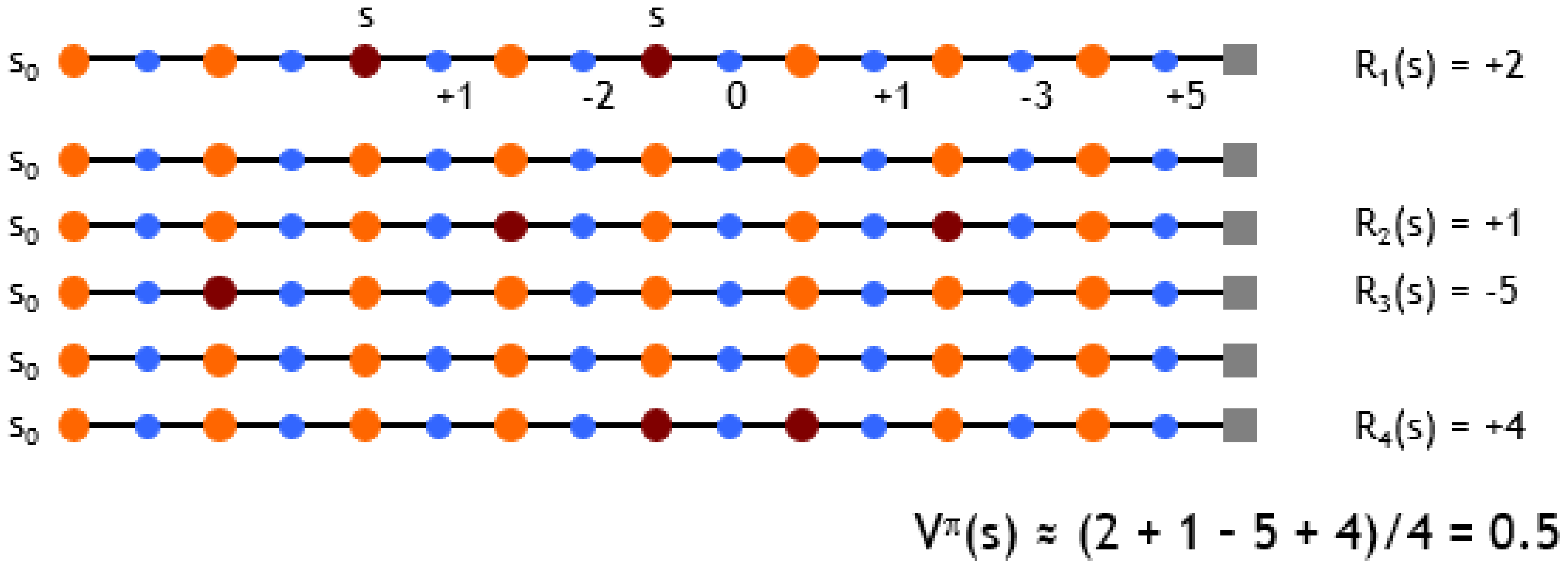
(b) For each state s appearing in the episode:

$R \leftarrow$ return following the first occurrence of s

Append R to $Returns(s)$

$V(s) \leftarrow \text{average}(Returns(s))$

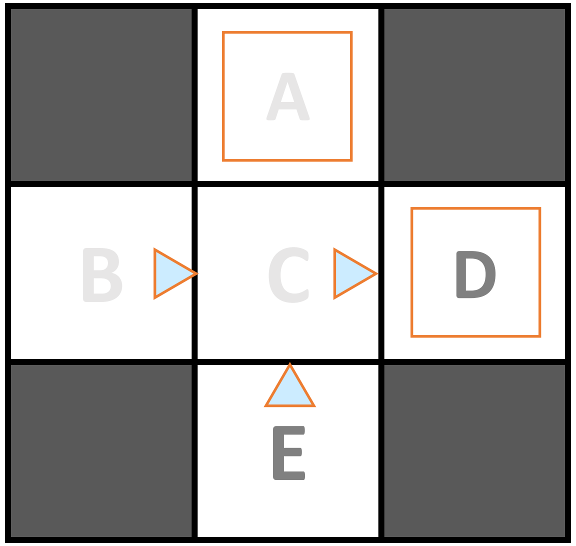
Ex-1:First-visit Monte-Carlo Policy Evaluation [estimate Vπ(s)]



Acknowledgements : This example is taken from the tutorial by Peter Bodík, RAD Lab, UC Berkeley

Ex-2: First-visit Monte-Carlo Policy Evaluation [estimate $V\pi(s)$]

Input Policy π



Assume: γ = 1

Observed Episodes (Training)

Episode 1

B, east, C, -1
C, east, D, -1
D, exit, , +10

Episode 2

B, east, C, -1
C, east, D, -1
D, exit, , +10

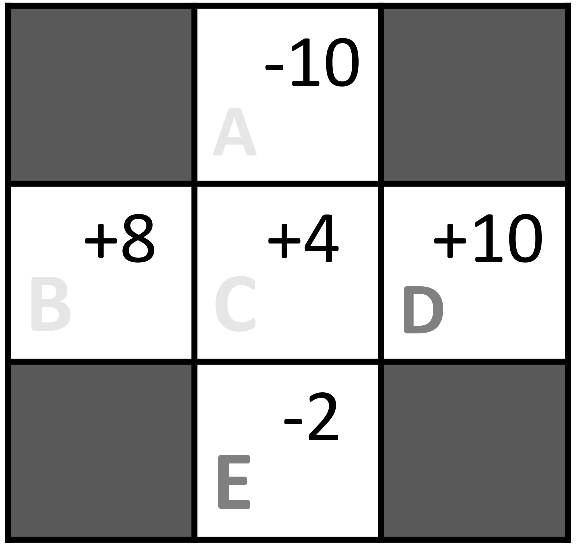
Episode 3

E, north, C, -1
C, east, D, -1
D, exit, , +10

Episode 4

E, north, C, -1
C, east, A, -1
A, exit, , -10

Output Values

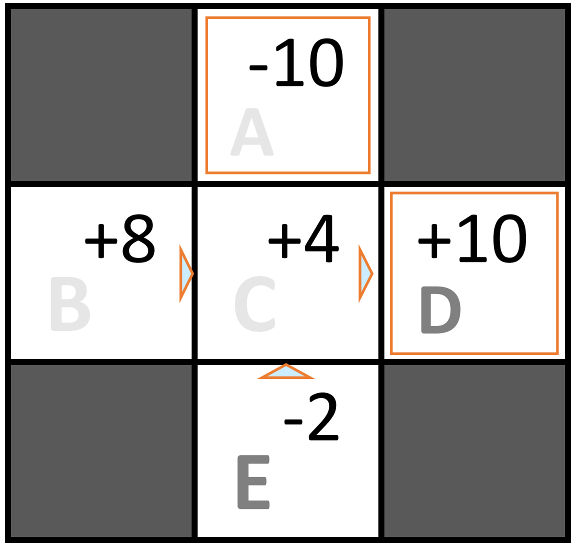


Problems with MC Evaluation

- What's good about direct evaluation?
  - It's easy to understand
  - It doesn't require any knowledge of the underlying model
  - It converges to the true expected values

- What bad about it?
  - It wastes information about transition probabilities
  - Each state must be learned separately
  - So, it takes a long time to learn

Output Values



Think: If B and E both go to C with the same probability, how can their values be different?

Must explore!

- Hard policy (insufficient): $\pi(s) = a\,,\ \ \pi\!: S \rightarrow \mathcal{A}$
- Soft policy: $\pi(a|s) = [0,1],\ \ \pi\!: S{\times}\mathcal{A} \rightarrow p$
  - At the beginning $\forall a,\ \pi(a|s) > 0$ to allow exploration
  - Gradually shift towards a deterministic policy
- For instance: select a random action with probability ε
  - $\forall a \neq A^*, \pi(s,a) = \frac{\varepsilon}{|\mathcal{A}(s)|}$
  - Else select the greedy action: $\pi(s, A^*) = 1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}(s)|}$

ε-greedy MC control

On-policy first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

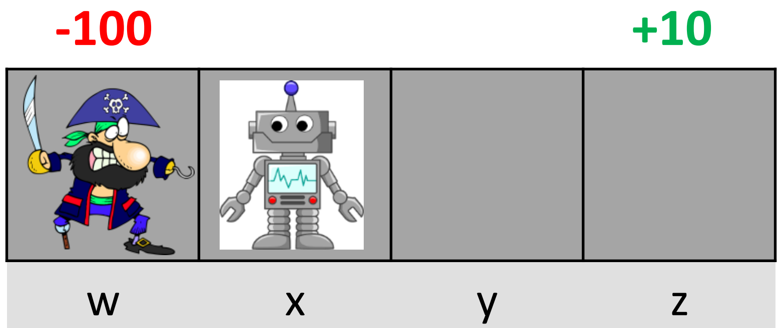
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

Repeat forever:
  (a) Generate an episode using π
  (b) For each pair s, a appearing in the episode:
    $G \leftarrow$ the return that follows the first occurrence of s, a
    Append G to $Returns(s,a)$
    $Q(s,a) \leftarrow \text{average}(Returns(s,a))$
  (c) For each s in the episode:
    $A^* \leftarrow \arg\max_a Q(s,a)$ (with ties broken arbitrarily)
    For all $a \in \mathcal{A}(s)$:

$$\pi(a|s) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(s)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq A^* \end{cases}$$

MC control - example

$\gamma = 0.9$



- $Q =$

5	4,3	2,1	0
w	x	y	z

- $\overline{Returns} =$

-	-,-	-,-	-
w	x	y	z

- $\pi(a|s) = (1 - \varepsilon) \cdot$
  - $\varepsilon \cdot$ Random

exit			exit
w	x	y	z

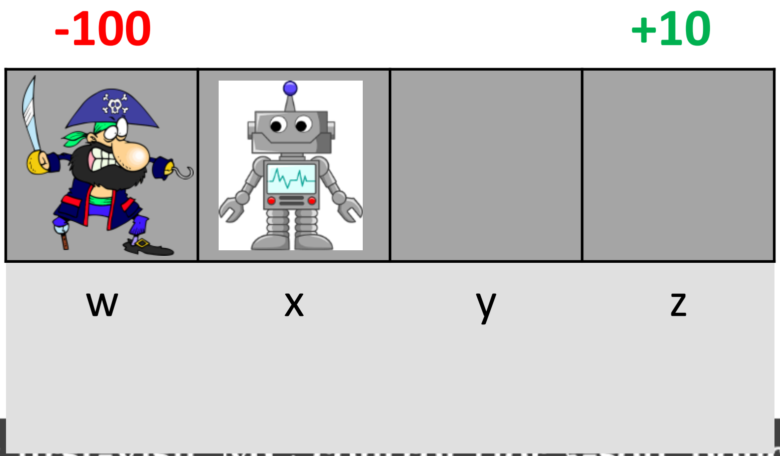
On-policy first-visit MC control (for ε-soft policies),

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

Repeat forever:
  (a) Generate an episode using π
  (b) For each pair s, a appearing in the episode:
    $G \leftarrow$ the return that follows the first occurrence of s, a
    Append G to $Returns(s,a)$
    $Q(s,a) \leftarrow \text{average}(Returns(s,a))$
  (c) For each s in the episode:
    $A^* \leftarrow \arg\max_a Q(s,a)$ (with ti
    For all $a \in \mathcal{A}(s)$:
$$\pi(a|s) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(s)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq A^* \end{cases}$$

MC control - example

$\gamma = 0.9$



- $Q =$

5	4,3	2,1	0
w	x	y	z

- $\overline{Returns} =$

-	-,-	-,-	-
w	x	y	z

- $\pi(a|s) = (1-\varepsilon) \cdot$

exit			exit
w	x	y	z

  - $\varepsilon \cdot$ Random

- $\tau = x, \leftarrow, 0, w, exit, -100$

On-policy first-visit MC control (for ε-soft policies),

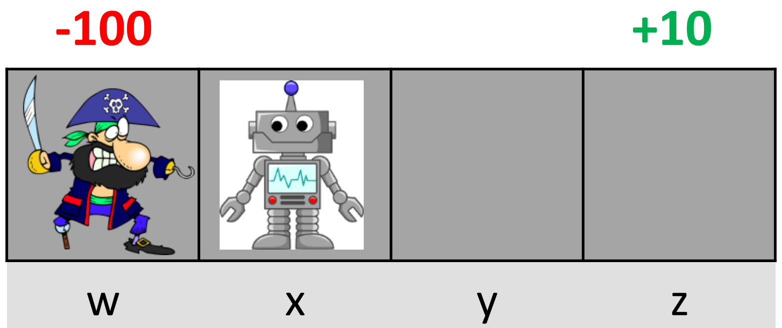
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

Repeat forever:
  (a) Generate an episode using π
  (b) For each pair s, a appearing in the episode:
    $G \leftarrow$ the return that follows the first occurrence of s, a
    Append G to $Returns(s,a)$
    $Q(s,a) \leftarrow \text{average}(Returns(s,a))$
  (c) For each s in the episode:
    $A^* \leftarrow \arg\max_a Q(s,a)$ (with ti
    For all $a \in \mathcal{A}(s)$:

$$\pi(a|s) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(s)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq A^* \end{cases}$$

MC control - example

$\gamma = 0.9$



- $Q =$

w	x	y	z
5	4,3	2,1	0

- $\overline{Returns} =$

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

- $\pi(a|s) = (1 - \varepsilon) \cdot$

w	x	y	z
exit			exit

  - $\varepsilon \cdot$ Random

- $\tau = x, \leftarrow, 0, w, exit, -100$

On-policy first-visit MC control (for ε-soft policies),

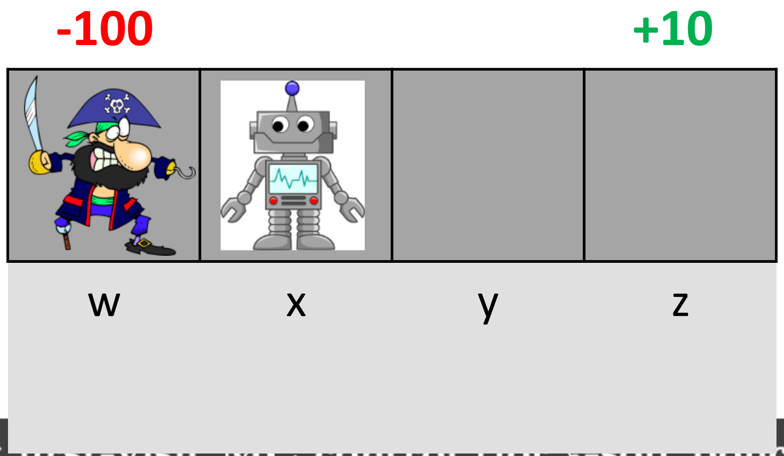
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

Repeat forever:
  (a) Generate an episode using π
  (b) For each pair s, a appearing in the episode:
    $G \leftarrow$ the return that follows the first occurrence of s, a
    Append G to $Returns(s,a)$
    $Q(s,a) \leftarrow \text{average}(Returns(s,a))$
  (c) For each s in the episode:
    $A^* \leftarrow \arg\max_a Q(s,a)$ (with ti
    For all $a \in \mathcal{A}(s)$:

$$\pi(a|s) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(s)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq A^* \end{cases}$$

$\gamma = 0.9$

MC control - example



- $Q =$

-100	-90,3	2,1	0
w	x	y	z

- $\overline{Returns} =$

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

- $\pi(a|s) = (1-\varepsilon) \cdot$

exit			exit
w	x	y	z

  - $\varepsilon \cdot$ Random

- $\tau = x, \leftarrow, 0, w, exit, -100$

On-policy first-visit MC control (for ε-soft policies),

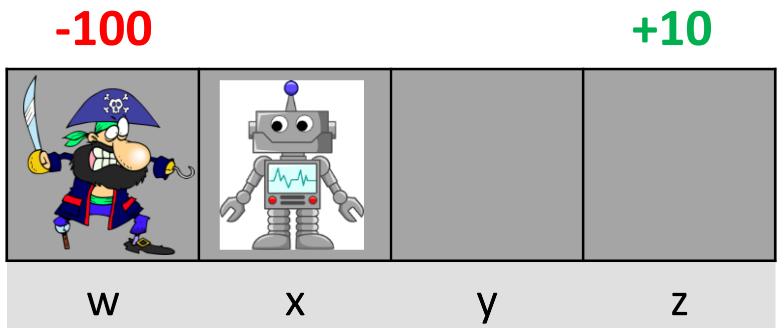
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

Repeat forever:
  (a) Generate an episode using π
  (b) For each pair s, a appearing in the episode:
    $G \leftarrow$ the return that follows the first occurrence of s, a
    Append G to $Returns(s,a)$
    $Q(s,a) \leftarrow \text{average}(Returns(s,a))$
  (c) For each s in the episode:
    $A^* \leftarrow \arg\max_a Q(s,a)$ (with ti
    For all $a \in \mathcal{A}(s)$:

$$\pi(a|s) \leftarrow \begin{cases} 1-\varepsilon+\varepsilon/|\mathcal{A}(s)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq A^* \end{cases}$$

MC control - example

$\gamma = 0.9$



- $Q =$

-100	-90,3	2,1	0
w	x	y	z

- $\overline{Returns} =$

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

- $\pi(a|s) = (1-\varepsilon) \cdot$

exit			exit
w	x	y	z

  - $\varepsilon \cdot$ Random

- $\tau = x, \leftarrow, 0, w, exit, -100$
- $A^* = [\rightarrow, exit]$

On-policy ... **policies),**

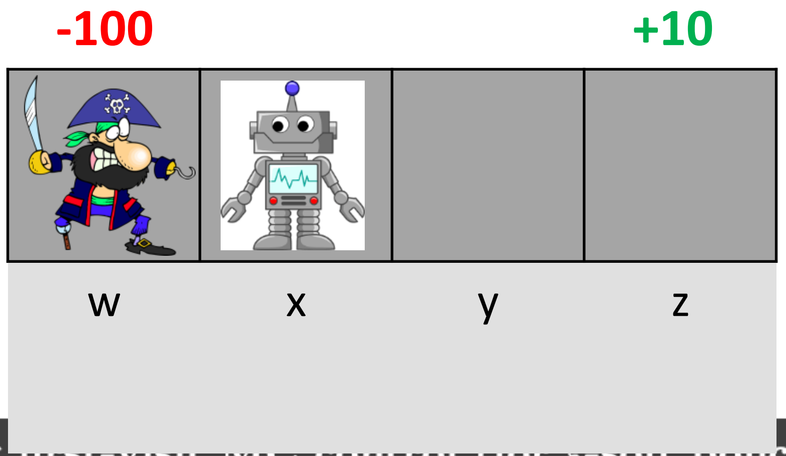
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

Repeat forever:
  (a) Generate an episode using π
  (b) For each pair s, a appearing in the episode:
    $G \leftarrow$ the return that follows the first occurrence of s, a
    Append G to $Returns(s,a)$
    $Q(s,a) \leftarrow \text{average}(Returns(s,a))$
  (c) For each s in the episode:
    $A^* \leftarrow \arg\max_a Q(s,a)$ (with ti
    For all $a \in \mathcal{A}(s)$:

$$\pi(a|s) \leftarrow \begin{cases} 1-\varepsilon+\varepsilon/|\mathcal{A}(s)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq A^* \end{cases}$$

MC control - example

$\gamma = 0.9$



- $Q =$

-100	-90,3	2,1	0
w	x	y	z

- $\overline{Returns} =$

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

- $\pi(a|s) = (1-\varepsilon) \cdot$

exit			exit
w	x	y	z

  - $\varepsilon \cdot$ Random

- $\tau = x, \leftarrow, 0, w, exit, -100$
- $A^* = [\rightarrow, exit]$

On-policy first-visit MC control (for ε-soft policies),

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

Repeat forever:
  (a) Generate an episode using π
  (b) For each pair s, a appearing in the episode:
    $G \leftarrow$ the return that follows the first occurrence of s, a
    Append G to $Returns(s,a)$
    $Q(s,a) \leftarrow \text{average}(Returns(s,a))$
  (c) For each s in the episode:
    $A^* \leftarrow \arg\max_a Q(s,a)$ (with ti
    For all $a \in \mathcal{A}(s)$:

$$\pi(a|s) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(s)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq A^* \end{cases}$$

MC control - example

- $Q =$

w	x	y	z
-100	-90,3	2,1	0

- $\overline{Returns} =$

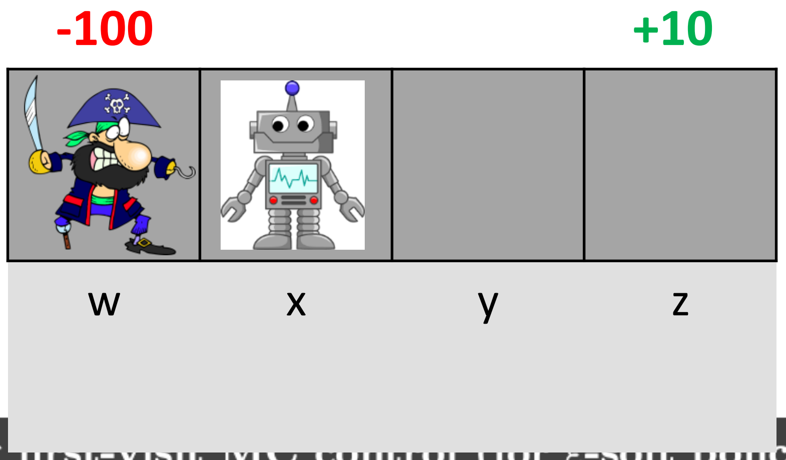
w	x	y	z
-100	-90,0	-,-	-

- $\pi(a|s) = (1 - \varepsilon) \cdot$
  - $\varepsilon \cdot$ Random

w	x	y	z
exit			exit

- $\tau = x, \rightarrow, 0, y, \leftarrow, 0, x, \leftarrow, 0, exit, -100$

$\gamma = 0.9$



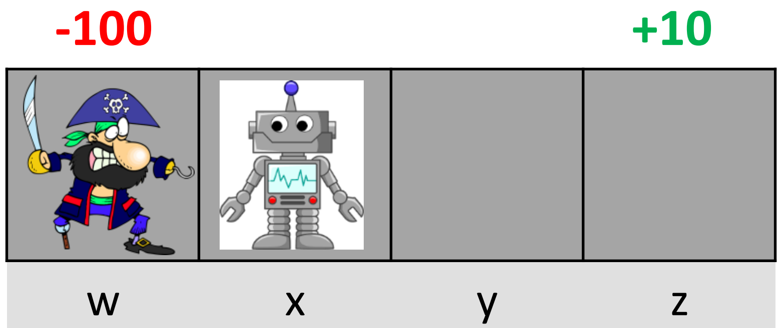
On-policy first-visit MC control (for ε-soft policies),

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

Repeat forever:
  (a) Generate an episode using π
  (b) For each pair s, a appearing in the episode:
    $G \leftarrow$ the return that follows the first occurrence of s, a
    Append G to $Returns(s,a)$
    $Q(s,a) \leftarrow \text{average}(Returns(s,a))$
  (c) For each s in the episode:
    $A^* \leftarrow \arg\max_a Q(s,a)$ (with ti
    For all $a \in \mathcal{A}(s)$:

$$\pi(a|s) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(s)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq A^* \end{cases}$$

$\gamma = 0.9$



MC control - example

- $Q =$

-100	-90,-72.9	-81,1	0
w	x	y	z

- $\overline{Returns} =$

-100	-90,-72.9	-81,-	-
w	x	y	z

- $\pi(a|s) = (1-\varepsilon) \cdot$

exit			exit
w	x	y	z

  - $\varepsilon \cdot$ Random

- $\tau = x, \rightarrow, 0, y, \leftarrow, 0, x, \leftarrow, 0, exit, -100$

On-policy first-visit MC control (for ε-soft policies),

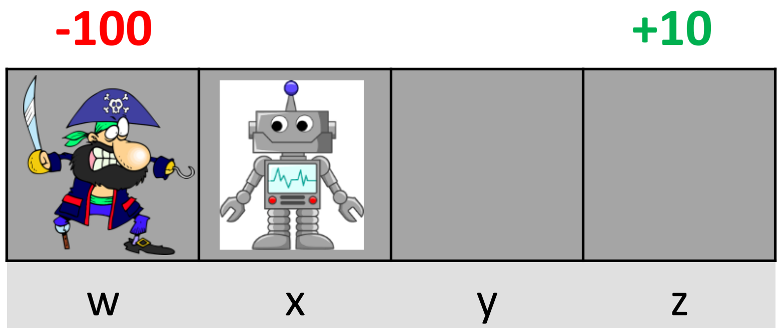
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

Repeat forever:
  (a) Generate an episode using π
  (b) For each pair s, a appearing in the episode:
    $G \leftarrow$ the return that follows the first occurrence of s, a
    Append G to $Returns(s,a)$
    $Q(s,a) \leftarrow \text{average}(Returns(s,a))$
  (c) For each s in the episode:
    $A^* \leftarrow \arg\max_a Q(s,a)$ (with ti
    For all $a \in \mathcal{A}(s)$:

$$\pi(a|s) \leftarrow \begin{cases} 1-\varepsilon+\varepsilon/|\mathcal{A}(s)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq A^* \end{cases}$$

MC control - example

$\gamma = 0.9$



- $Q =$

w	x	y	z
-100	-90,-72.9	-81,1	0

- $\overline{Returns} =$

w	x	y	z
-100	-90,-72.9	-81,-	-

- $\pi(a|s) = (1-\varepsilon) \cdot$
  - $\varepsilon \cdot$ Random

w	x	y	z
exit			exit

- $\tau = x, \rightarrow, 0, y, \leftarrow, 0, x, \leftarrow, 0, exit, -100$
- $A^* = [\rightarrow, \rightarrow, exit]$

On-policy first-visit MC control (for ε-soft policies),

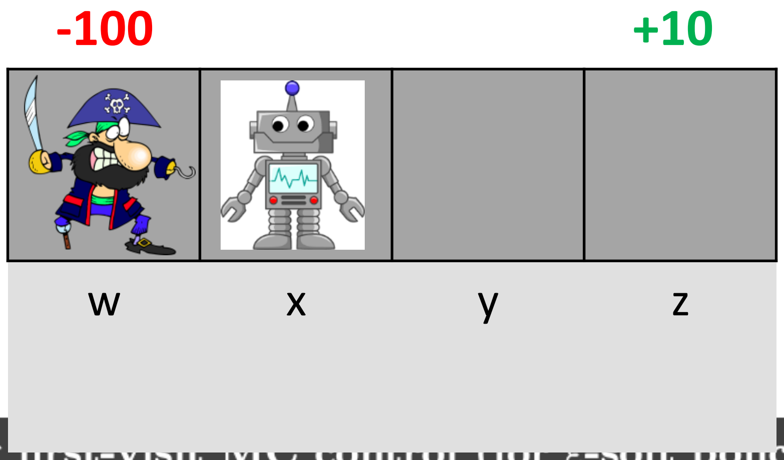
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

Repeat forever:
  (a) Generate an episode using π
  (b) For each pair s, a appearing in the episode:
    $G \leftarrow$ the return that follows the first occurrence of s, a
    Append G to $Returns(s,a)$
    $Q(s,a) \leftarrow \text{average}(Returns(s,a))$
  (c) For each s in the episode:
    $A^* \leftarrow \arg\max_a Q(s,a)$ (with ti
    For all $a \in \mathcal{A}(s)$:

$$\pi(a|s) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(s)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq A^* \end{cases}$$

MC control - example

$\gamma = 0.9$



- $Q =$

-100	-90,-72.9	-81,1	0
w	x	y	z

- $\overline{Returns} =$

-100	-90,-72.9	-81,-	-
w	x	y	z

- $\pi(a|s) = (1-\varepsilon) \cdot$

exit			exit
w	x	y	z

  - $\varepsilon \cdot$ Random

- $\tau = x, \rightarrow, 0, y, \leftarrow, 0, x, \leftarrow, 0, exit, -100$
- $A^* = [\rightarrow, \rightarrow, exit]$

On-policy first-visit MC control (for ε-soft policies),

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

Repeat forever:
  (a) Generate an episode using π
  (b) For each pair s, a appearing in the episode:
    $G \leftarrow$ the return that follows the first occurrence of s, a
    Append G to $Returns(s,a)$
    $Q(s,a) \leftarrow \text{average}(Returns(s,a))$
  (c) For each s in the episode:
    $A^* \leftarrow \arg\max_a Q(s,a)$ (with ti
    For all $a \in \mathcal{A}(s)$:

$$\pi(a|s) \leftarrow \begin{cases} 1-\varepsilon+\varepsilon/|\mathcal{A}(s)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq A^* \end{cases}$$

Quick Recap !
On-policy vs. Off-policy Learning

Required Readings

1. Chapter-3,4 of Introduction to Reinforcement Learning,2^{nd} Ed., Sutton & Barto

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Thank you