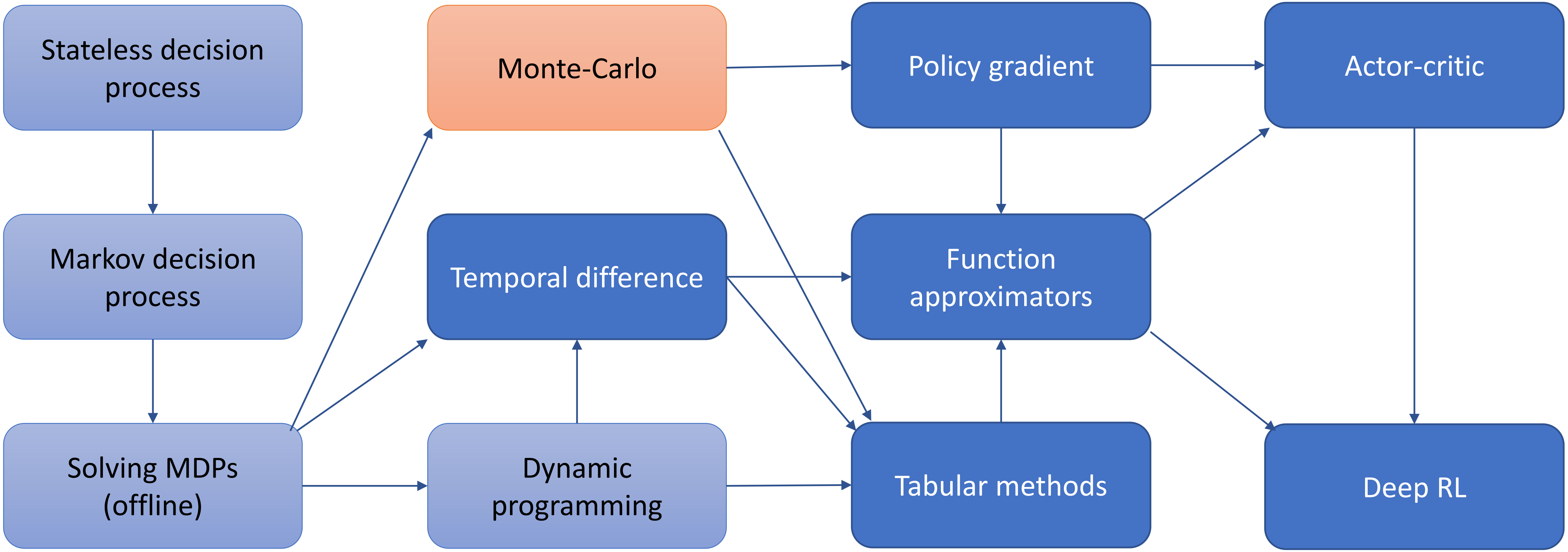
CSCE-642 Reinforcement Learning
Chapter 5: Monte Carlo Methods

Instructor: Guni Sharon

CSCE-689, Reinforcement Learning



Reinforcement Learning

- 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 $\pi^*(s)$

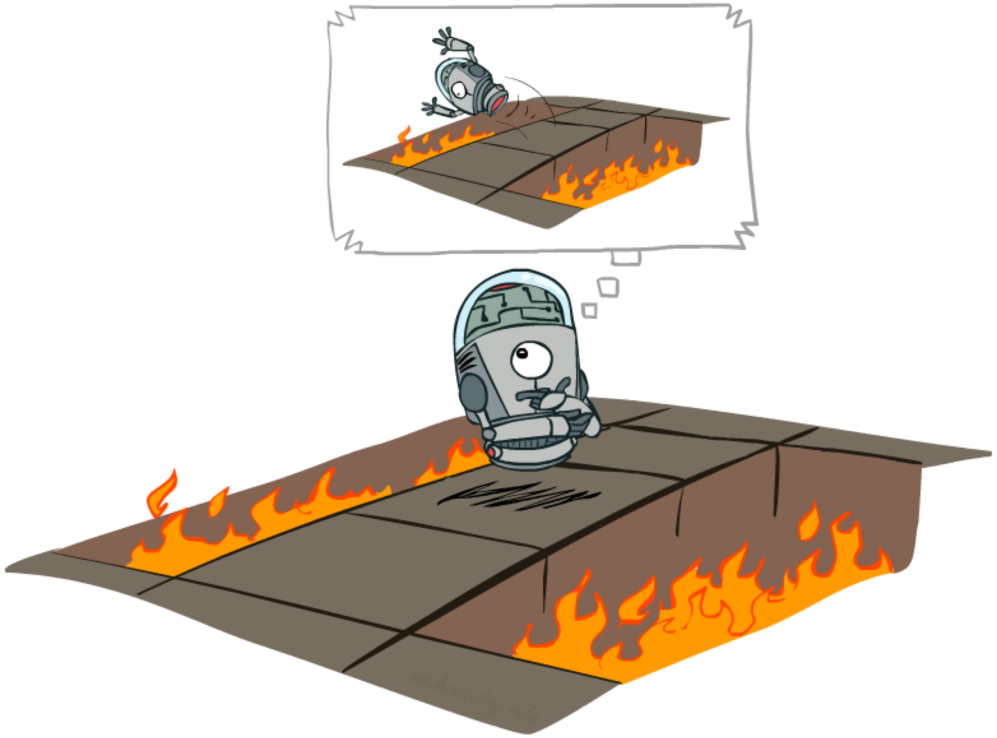
Cool

Warm

Overheated

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

Offline vs. Online (RL)



Offline Optimization

Online Learning

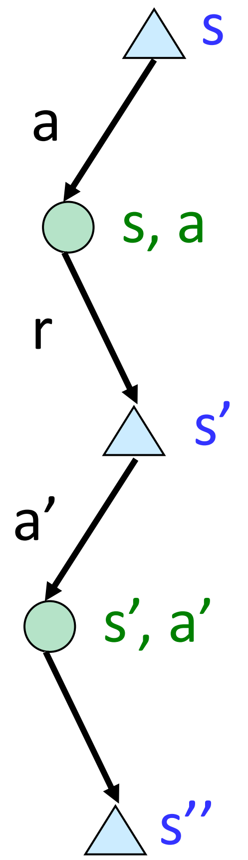
Monte-Carlo Methods

- 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

Value-based learning via Monte-Carlo

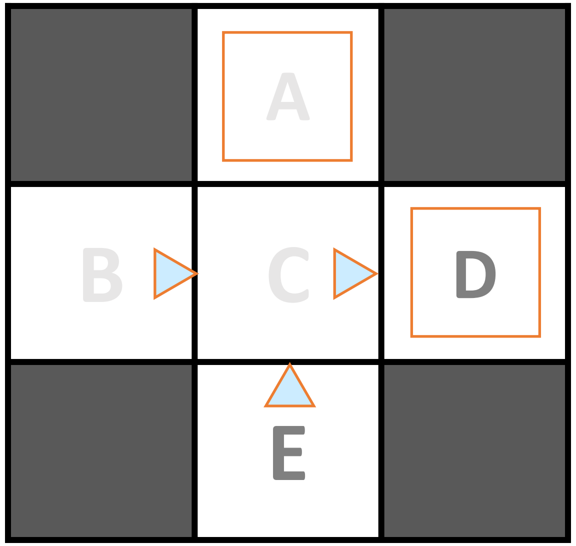
1. Learn an unbiased evaluation of q_π through sampling

- Experience world through episodes
  - $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$
- Update estimates for each transition
  - $Q(S, A) \leftarrow AVG_{episodes}(G_t | S_t = S, A_t = A)$
- Over time, values will converge to q_π



Example: MC Evaluation

Input Policy π



Assume: $\gamma = 1$

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

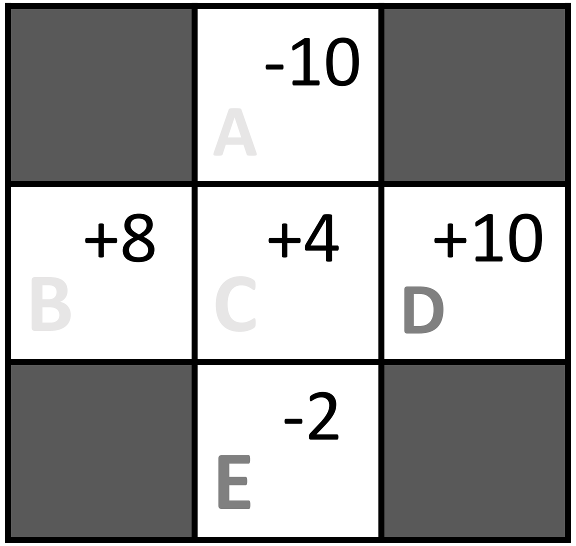
Episode 3

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

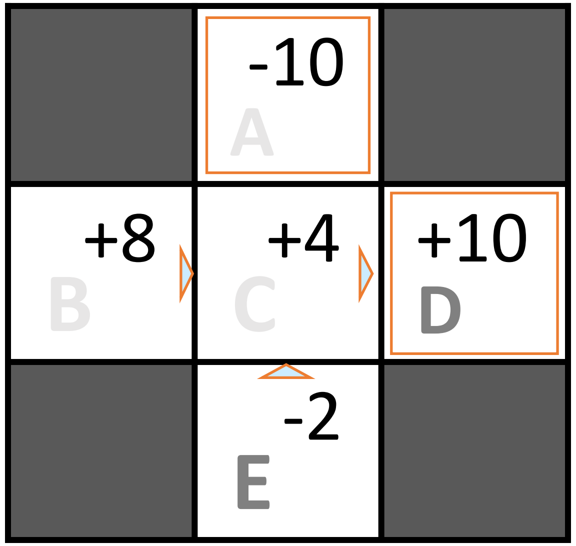


Problems with MC Evaluation

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

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

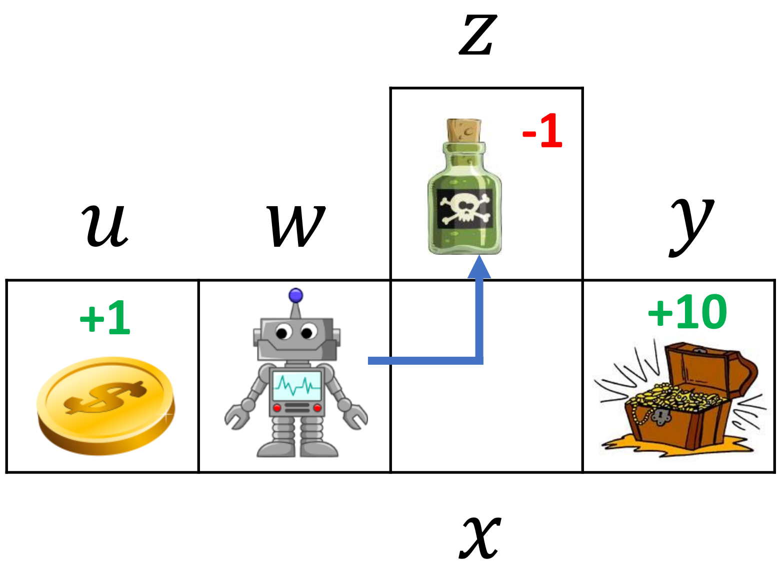
Output Values



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

Greedy MC

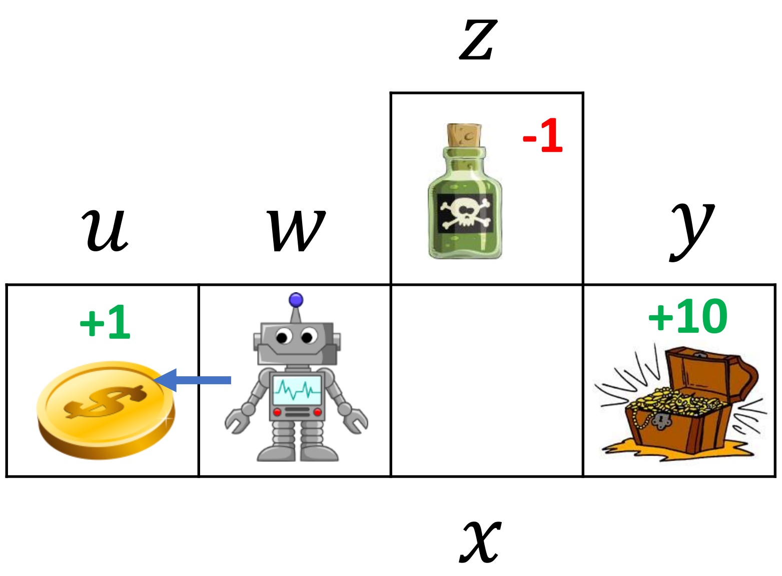
- $S = \{u, w, x, y, z\}$
- $A = \{N, E, S, W, exit\}$
- Reward:
  - $r(u, exit) = 1$
  - $r(z, exit) = -1$
  - $r(y, exit) = 10$



State	$\pi(s)$	$Q_\pi(N)$	$Q_\pi(E)$	$Q_\pi(S)$	$Q_\pi(W)$	$Q_\pi(exit)$
u	$exit$	NA	NA	NA	NA	0
w	E	0	$\cancel{0}$ -1	0	0	NA
x	N	$\cancel{0}$ -1	0	0	0	NA
y	$exit$	NA	NA	NA	NA	0
z	$exit$	NA	NA	NA	NA	$\cancel{0}$ -1

Greedy MC

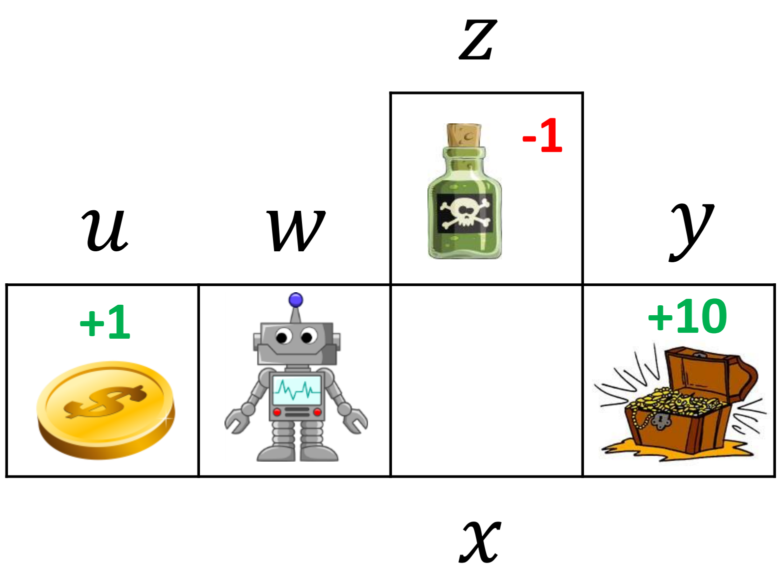
- $S = \{u, w, x, y, z\}$
- $A = \{N, E, S, W, exit\}$
- Reward:
  - $r(u, exit) = 1$
  - $r(z, exit) = -1$
  - $r(y, exit) = 10$



State	$\pi(s)$	$Q_\pi(N)$	$Q_\pi(E)$	$Q_\pi(S)$	$Q_\pi(W)$	$Q_\pi(exit)$
u	$exit$	NA	NA	NA	NA	$\cancel{0}$ 1
w	$\cancel{E}$ W	0	-1	0	$\cancel{0}$ 1	NA
x	$\cancel{N}$ E	-1	0	0	0	NA
y	$exit$	NA	NA	NA	NA	0
z	$exit$	NA	NA	NA	NA	-1

Greedy MC

- $S = \{u, w, x, y, z\}$
- $A = \{N, E, S, W, exit\}$
- Reward:
  - $r(u, exit) = 1$
  - $r(z, exit) = -1$
  - $r(y, exit) = 10$



State	$\pi(s)$	$Q_\pi(N)$	$Q_\pi(E)$	$Q_\pi(S)$	$Q_\pi(W)$	$Q_\pi(exit)$
u	$exit$	NA	NA	NA	NA	1
w	$\boxed{W}$	0	[illegible]	[illegible]	1	NA
x	$\boxed{E}$	-1	[illegible]	[illegible]	0	NA
y	$exit$	NA	NA	NA	NA	0
z	$exit$	NA	NA	NA	NA	-1

We converged on a local optimum!

Must explore!

- Hard policy (insufficient): $\pi: S \rightarrow \mathcal{A}$
- 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 $\pi \approx \pi_*$

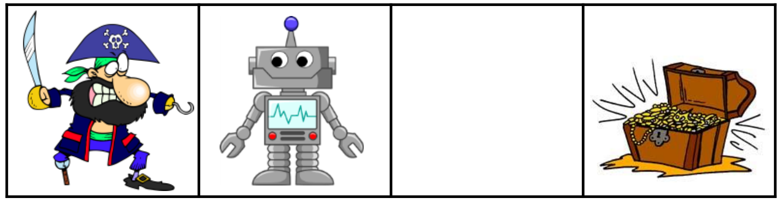
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s, a) \leftarrow$ arbitrary
  $Returns(s, a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

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 $a \in \mathcal{A}(s)$:

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

$\gamma = 0.9$

MC control - example



- $Q =$

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

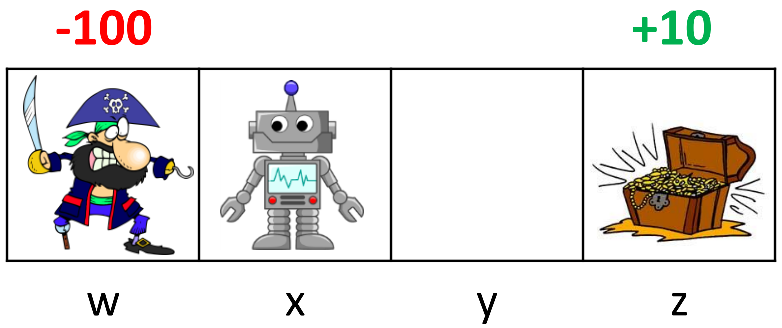
On-policy first-visit MC control (for ε-soft policies),

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



- $Q =$

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

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

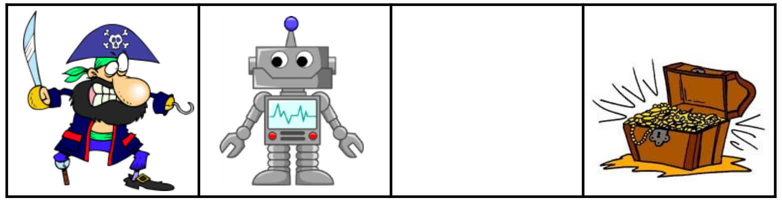
On-policy first-visit MC control (for ε-soft policies),

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

$\gamma = 0.9$

MC control - example



- $Q =$

5	4,3	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),

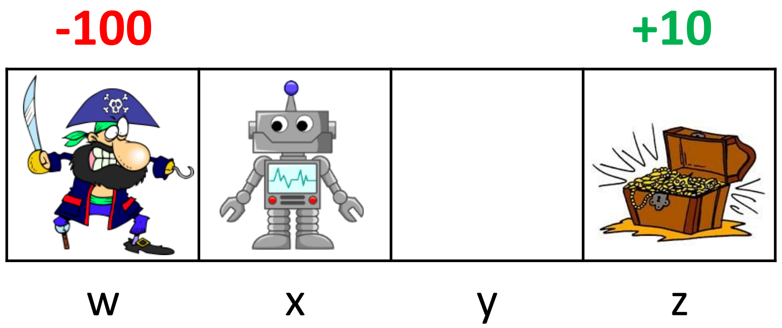
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

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

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

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

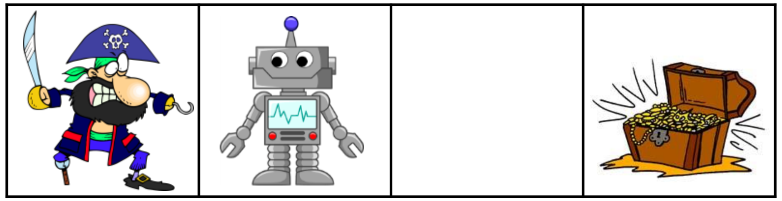
- $\pi(a|s) = (1-\varepsilon) \cdot$

exit	←	←	exit
w	x	y	z

  - $\varepsilon \cdot$ Random

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

-100 +10



w x y z

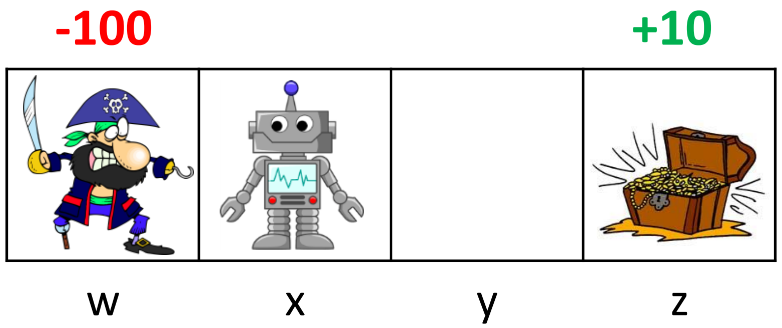
On-policy first-visit MC control (for ε-soft policies),

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 $a \in \mathcal{A}(s)$:

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

$\gamma = 0.9$



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

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

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

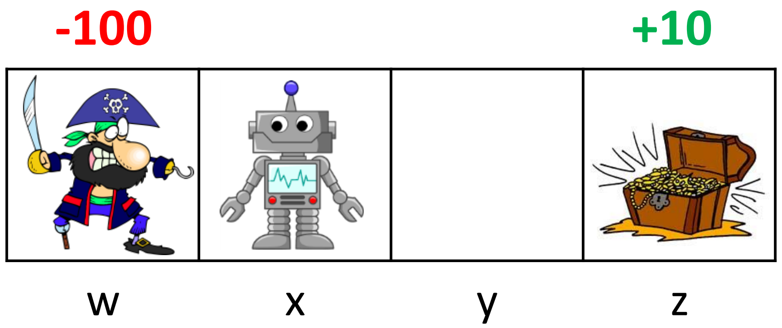
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

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 $a \in \mathcal{A}(s)$:

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

$\gamma = 0.9$

MC control - example



- $Q =$

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

- $\overline{Returns} =$

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

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

w	x	y	z
exit	→	←	exit

  - $\varepsilon \cdot$ Random

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

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

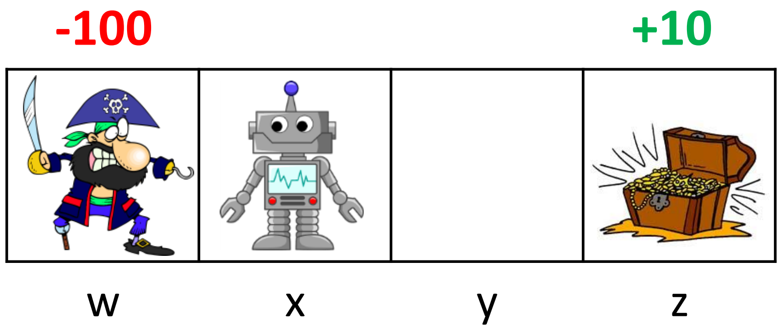
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

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 $a \in \mathcal{A}(s)$:

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

$\gamma = 0.9$

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),

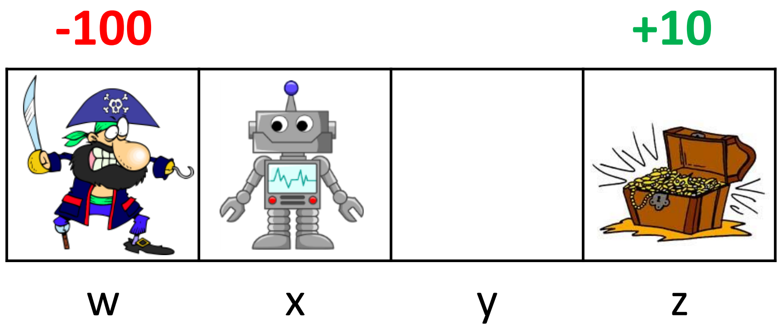
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $Returns(s,a) \leftarrow$ empty list
  $\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

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 $a \in \mathcal{A}(s)$:

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

$\gamma = 0.9$

MC control - example



- $Q =$

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

- $\overline{Returns} =$

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

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

w	x	y	z
exit	→	←	exit

  - $\varepsilon \cdot$ Random

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

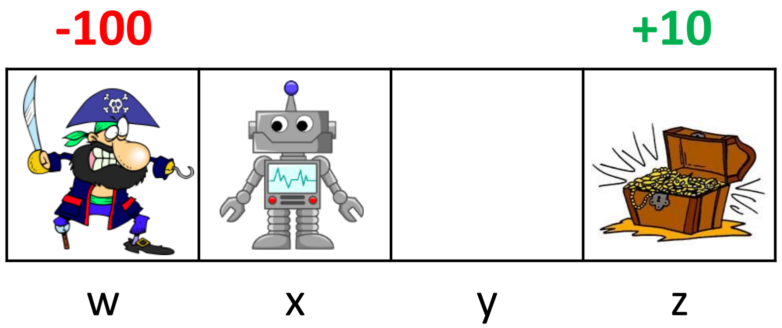
On-policy first-visit MC control (for ε-soft policies),

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 $a \in \mathcal{A}(s)$:

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

$\gamma = 0.9$



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

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

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

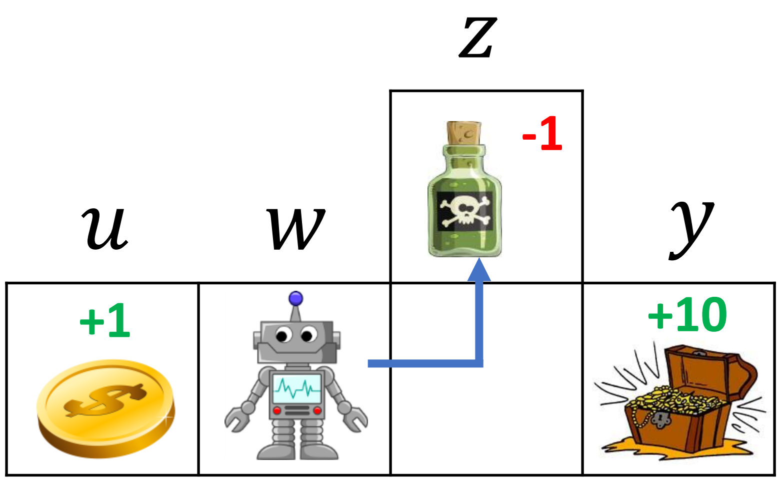
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
$Q(s,a) \leftarrow$ arbitrary
$Returns(s,a) \leftarrow$ empty list
$\pi(a|s) \leftarrow$ an arbitrary ε-soft policy

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 $a \in \mathcal{A}(s)$:

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

On-policy learning

- $\underbrace{Q_\pi(w, \rightarrow)}_{\text{Estimation}} = \underbrace{q_\pi(w, \rightarrow)}_{\text{True value}} = -1$ (correct!)
- $q^*(w, \rightarrow) = ?$
- $q_{\pi \neq b} = ?$
- Observation drawn from π are useful for evaluating q_π
- Once the policy is changed these observations are irrelevant
- This is not **sample efficient**!



State	$\pi(s)$	$Q_\pi(\uparrow)$	$Q_\pi(\rightarrow)$
u	$exit$	NA	NA
w	$\rightarrow$	0	-1
x	$\uparrow$	-1	0
y	$exit$	NA	NA
z	$exit$	NA	NA

Off-policy learning

- We would like to use observations drawn from some policy b to evaluate q_π where $\pi \neq b$, specifically, we want to evaluate q_{π^*}
- We strive for full utilization of previous experience
- Off-policy learning allows us to optimize a ***target policy*** while following another ***behavior policy***
- **Pros:** sample efficient
- **Cons:** higher variance in value estimations

Off-policy learning conditions

- **Objective**: use episodes from b to estimate values for π
- For off-policy learning we must assume ***coverage***
  - $\forall s, a,\ \pi(a|s) > 0 \Longrightarrow b(a|s) > 0$
- If this is true, then by running b repeatedly we will eventually discover all possible trajectories for π
- If coverage is violated for some s, a, then no inference is possible regarding that state-action value

Trajectory probability

- An agent following policy b sampled the following trajectory
  - $\tau = \{S_0, A_0, R_1, S_1, A_1, R_2, \dots, A_{T-1}, R_T, S_T\}$
- What is the probability of sampling this trajectory?
  - $\Pr\{\tau|b\} = b(A_0|S_0)p(S_1|S_0, A_0)b(A_1|S_1)p(S_2|S_1, A_1) \dots b(A_{T-1}|S_{T-1})p(S_T|S_{T-1}, A_{T-1})$
  - $= \prod_{k=0}^{T-1} b(A_K|S_K)p(S_{K+1}|S_K, A_K)$
- Assume MC control, how should $Q_b(S_t, A_t)$ be updated?
  - $G_t = \sum_{k=t+1}^{T} \gamma^{k-t+1} R_k$
  - Append G to $Returns(S_t, A_t)$ with weight $\Pr_t\{\tau|b\}$
  - $Q_b(S_t, A_t) = \text{Weighted\_AVG}\big(Returns(S_t, A_t)\big)$

$$\mathbb{E}[X \sim p] = \sum_x p(x)x$$

- Computing the weighted average based on the sample probability would reduce variance (eliminate impact of noisy sampling)
- But the model, $p(S_{K+1}|S_K, A_K)$, is unknown! So can't compute $\Pr\{\tau|b\}$
- Can we say anything about $\Pr\{\tau|\pi \neq b\}$?

Importance sampling

- Given a trajectory τ drawn by running b
- We can **define** (not compute) the probability $\Pr\{\tau|b\}$
- We can also **define** $\Pr\{\tau|\pi\}$
- Define the ***importance sampling ratio*** as: $\rho_t = \frac{\Pr\{\tau_t|\pi\}}{\Pr\{\tau_t|b\}}$
- Can we **compute** ρ without a model, $p(S_{K+1}|S_K, A_K)$?
- $\rho_t = \frac{\prod_{k=t}^{T-1} \pi(A_K|S_K)\cancel{p(S_{K+1}|S_K,A_K)}}{\prod_{k=t}^{T-1} b(A_K|S_K)\cancel{p(S_{K+1}|S_K,A_K)}} = \prod_{k=t}^{T-1} \frac{\pi(A_K|S_K)}{b(A_K|S_K)}$ **YES!**

Importance sampling

- **By definition**: $\mathbb{E}_{\tau\sim b}[G_t|S_t = s] = v_b(s)$
- Importance sampling allows us to compute an unbiased estimate of $v_\pi(s)$ by running b
- **Claim**: $\mathbb{E}_{\tau\sim b}[\rho_t G_t|S_t = s] = v_\pi(s)$
- We set $v_\pi(s)$ to be a weighted average of observed returns (weighted by the importance ratio)
- Assume visiting state s over M episodes using policy b
  - s is first visited during time step t^m during each episode, $\mathrm{m} \in M$
- $v_\pi(s) = \frac{\sum_{\mathrm{m}\in M} \rho_{t^m} G_{t^m}}{M}$

Importance sampling: proof

- We would like to estimate $\mathbb{E}[X]$ in our case when $X = v_\pi(s)$
- By definition: $\mathbb{E}[X] = \sum_x \Pr_{\mathrm{X}}(x)x$
  - for the continuous case replace the sum with an integral
- If we don't know $\Pr(x)$ we can take a sample-based approach
- $\mathbb{E}[X] = \sum_x \Pr_{\mathrm{X}}(x)x = \frac{\sum_{i=0}^{n} x_i}{n}$
  - This is an unbiased estimation because the samples are coming from the same distribution that defines $\Pr_X$
- But what if our samples come from a different distribution $\Pr_Y$?

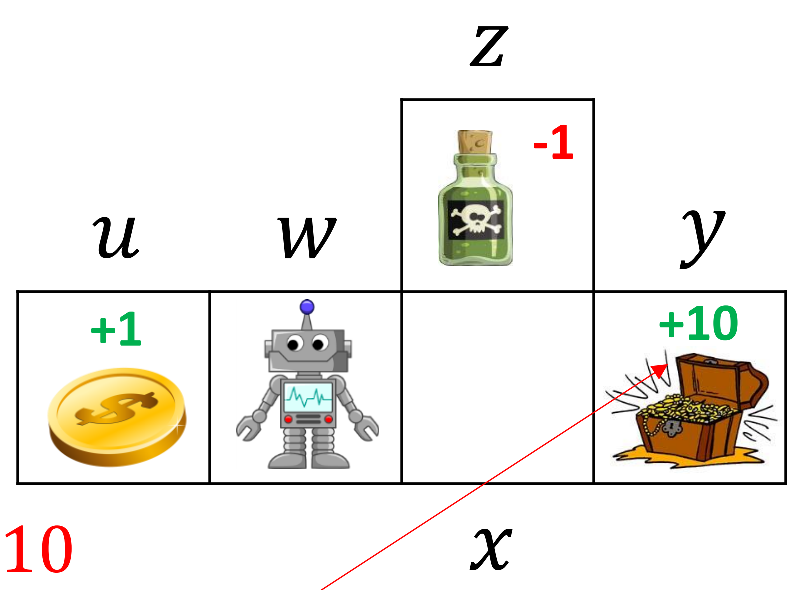
Importance sampling: proof

- Assume we know: $\mathbb{E}[Y] = \sum_y \Pr_{Y}(y) y = \frac{\sum_{i=0}^{n} y_i}{n}$
- Can we use this to compute $\mathbb{E}[X]$?
- Yes! if we assume that all possible values of X exist in Y (*converge*)
- $\mathbb{E}[X] = \sum_x \Pr_X(x) x = \sum_y \Pr_X(y) y$
- $= \sum_y \frac{\Pr_X(y)}{\Pr_Y(y)} \Pr_Y(y) y$
- $= \frac{\Pr_X(y)}{\Pr_Y(y)} \cdot \frac{\sum_{i=0}^{n} y_i}{n}$

Importance ratio

(ordinary) Importance sampling - example

- $b(s|a) = \begin{cases} \rightarrow, p(0.5) \\ \leftarrow, p(0.5) \end{cases}$
- $\pi(s|a) = \begin{cases} \rightarrow, p(0.99) \\ \leftarrow, p(0.01) \end{cases}$
- $\tau_1 = \{w, \rightarrow, 0, x, \rightarrow, 0, y, exit, 10\}$
- $v_\pi(w) = \frac{\sum_{m \in M} \rho_{t^m} G_t^m}{M} = \frac{0.99}{0.5} * \frac{0.99}{0.5} * 10 = 3.96 * 10$
- $\tau_2 = \{w, \leftarrow, 0, u, exit, 1\}$
- $v_\pi(w) = \frac{\sum_{m \in M} \rho_{t^m} G_t^m}{M} = \frac{3.96*10+0.02*1}{2}$

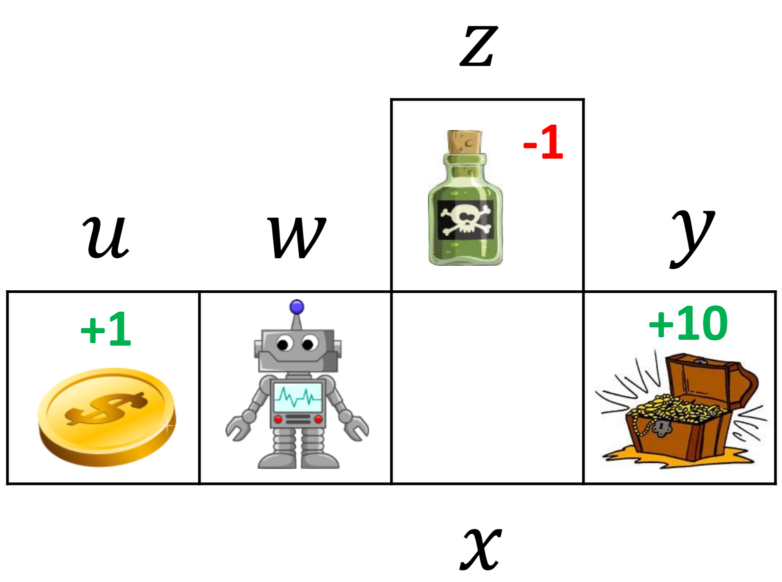


39.6 ? ?
Ordinary Importance sampling is unbiased yet high variance

Weighted importance sampling

- $v_\pi(s) = \frac{\sum_{m \in M}[\rho_{t^m} G_t^m]}{\sum_{m \in M} \rho_{t^m}}$
- $b(s|a) = \begin{cases} \rightarrow, p(0.5) \\ \leftarrow, p(0.5) \end{cases}$
- $\pi(s|a) = \begin{cases} \rightarrow, p(0.99) \\ \leftarrow, p(0.01) \end{cases}$
- $\tau_1 = \{w, \rightarrow, 0, x, \rightarrow, 0, y, exit, 10\}$
- $v_\pi(w) = \frac{\sum_{m \in M}[\rho_{t^m} G_t^m]}{\sum_{m \in M} \rho_{t^m}} = \frac{3.96*10}{3.96}$
- $\tau_2 = \{w, \leftarrow, 0, u, exit, 1\}$
- $v_\pi(w) = \frac{\sum_{m \in M}[\rho_{t^m} G_t^m]}{\sum_{m \in M} \rho_{t^m}} = \frac{3.96*10+0.02*1}{3.98}$

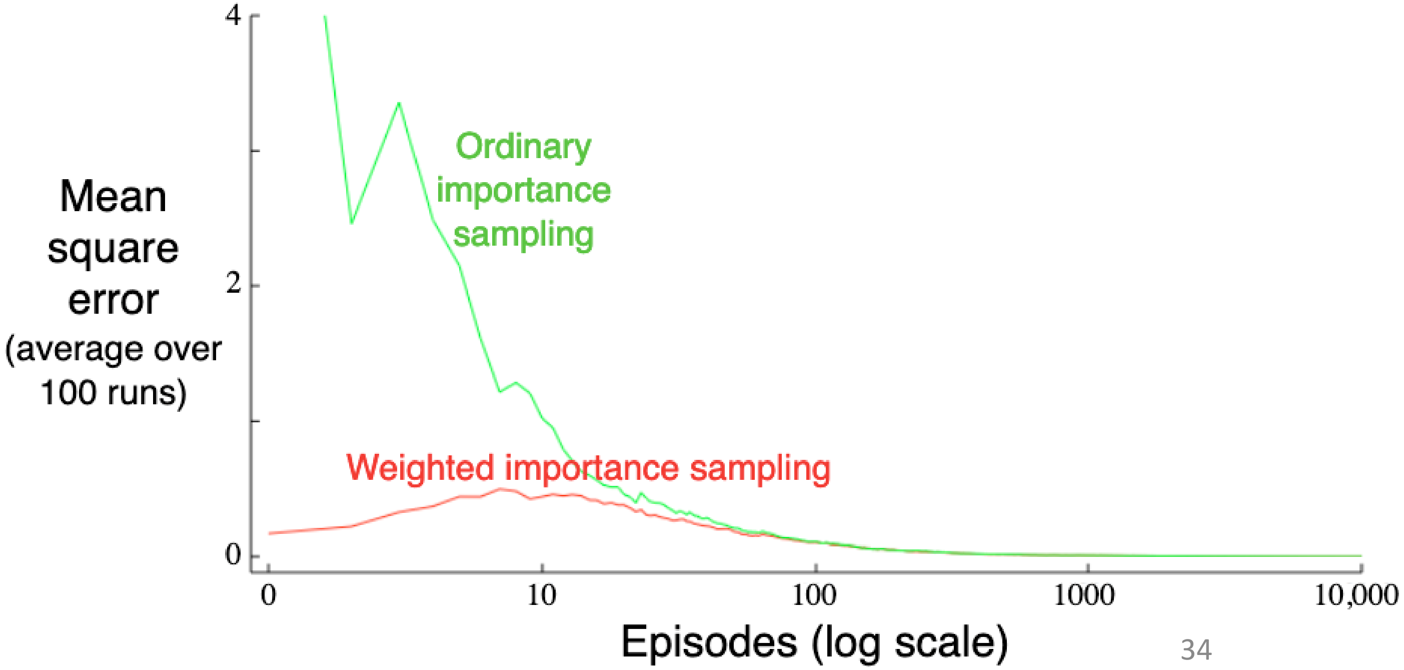
Trick: normalize by the sum of importance ratios



Ordinary Importance sampling is unbiased while the weighted version is biased (initially). Ordinary Importance sampling results in high variance while the weighted version has a bounded variance

Ordinary vs weighted importance sampling

- Estimating a black-jack state
- Target policy: hit on 19 or below
- Behavior policy: random (uniform)
- Both approaches converge to the true value
- weighted importance sampling is much better initially



MC control + importance sampling

Accumulated reward after step t

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $\boxed{G \leftarrow 0}$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t,A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + importance sampling

1 over Joint probability for observed actions from following b after step t. This equals the IS ratio here because the target policy is deterministic.

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s, a) \leftarrow$ arbitrary
    $C(s, a) \leftarrow 0$
    $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
    $b \leftarrow$ any soft policy
    Generate an episode using b:
        $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
    $G \leftarrow 0$
    $\boxed{W \leftarrow 1}$
    For $t = T-1, T-2, \ldots$ down to 0:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit For loop
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + importance sampling

Going back in time

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s, a) \leftarrow$ arbitrary
    $C(s, a) \leftarrow 0$
    $\pi(s) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
    $b \leftarrow$ any soft policy
    Generate an episode using b:
        $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
    $G \leftarrow 0$
    $W \leftarrow 1$
    $\boxed{\text{For } t = T-1, T-2, \ldots \text{ down to } 0:}$
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit For loop
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + importance sampling

Discount future rewards and add immediate reward

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
$Q(s,a) \leftarrow$ arbitrary
$C(s,a) \leftarrow 0$
$\pi(s) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
$b \leftarrow$ any soft policy
Generate an episode using b:
$S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
$G \leftarrow 0$
$W \leftarrow 1$
For $t = T-1, T-2, \ldots$ down to 0:
$\boxed{G \leftarrow \gamma G + R_{t+1}}$
$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)
If $A_t \neq \pi(S_t)$ then exit For loop
$W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + importance sampling

Cumulative sum of IS weights affiliated with S_t, A_t (for weighted IS)

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:

$Q(s,a) \leftarrow$ arbitrary

$C(s,a) \leftarrow 0$

$\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:

$b \leftarrow$ any soft policy

Generate an episode using b:

$S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$

$G \leftarrow 0$

$W \leftarrow 1$

For $t = T-1, T-2, \ldots$ down to 0:

$G \leftarrow \gamma G + R_{t+1}$

$\boxed{C(S_t, A_t) \leftarrow C(S_t, A_t) + W}$

$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$

$\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

If $A_t \neq \pi(S_t)$ then exit For loop

$W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + importance sampling

Incremental update of Q values (waited moving average)

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:

$Q(s, a) \leftarrow$ arbitrary

$C(s, a) \leftarrow 0$

$\pi(s) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:

$b \leftarrow$ any soft policy

Generate an episode using b:

$S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$

$G \leftarrow 0$

$W \leftarrow 1$

For $t = T-1, T-2, \ldots$ down to 0:

$G \leftarrow \gamma G + R_{t+1}$

$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$

$\boxed{Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]}$

$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)

If $A_t \neq \pi(S_t)$ then exit For loop

$W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + importance sampling

Update target policy (greedy)

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \arg\max_a Q(S_t,a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t,A_t) \leftarrow C(S_t,A_t) + W$
    $Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \frac{W}{C(S_t,A_t)}[G - Q(S_t,A_t)]$
    $\boxed{\pi(S_t) \leftarrow \arg\max_a Q(S_t,a) \quad \text{(with ties broken consistently)}}$
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + importance sampling

Sinse π is deterministic, once we diverge from it all IS weights of earlier actions will be 0

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
- $Q(s,a) \leftarrow$ arbitrary
- $C(s,a) \leftarrow 0$
- $\pi(s) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
- $b \leftarrow$ any soft policy
- Generate an episode using b:
  - $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
- $G \leftarrow 0$
- $W \leftarrow 1$
- For $t = T-1, T-2, \ldots$ down to 0:
  - $G \leftarrow \gamma G + R_{t+1}$
  - $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
  - $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
  - $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)
  - $\boxed{\text{If } A_t \neq \pi(S_t) \text{ then exit For loop}}$
  - $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + importance sampling

Update the joint prob by multiplying by ρ_t. Notice that $\pi(S_t) = 1$ in this example (deterministic target policy)

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:

$\quad Q(s, a) \leftarrow$ arbitrary

$\quad C(s, a) \leftarrow 0$

$\quad \pi(s) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:

$\quad b \leftarrow$ any soft policy

$\quad$ Generate an episode using b:

$\quad\quad S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$

$\quad G \leftarrow 0$

$\quad W \leftarrow 1$

$\quad$ For $t = T-1, T-2, \ldots$ down to 0:

$\quad\quad G \leftarrow \gamma G + R_{t+1}$

$\quad\quad C(S_t, A_t) \leftarrow C(S_t, A_t) + W$

$\quad\quad Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$

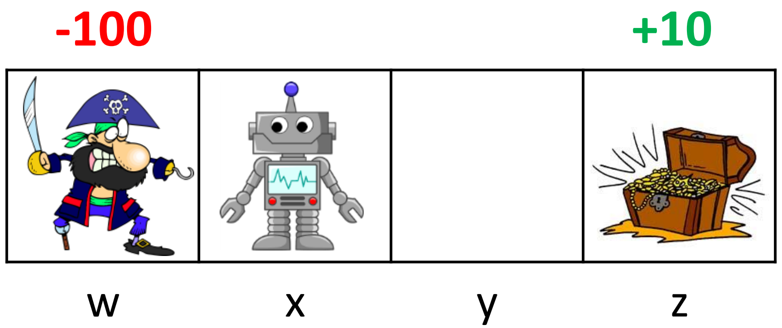
$\quad\quad \pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)

$\quad\quad$ If $A_t \neq \pi(S_t)$ then exit For loop

$\quad\quad \boxed{W \leftarrow W \frac{1}{b(A_t|S_t)}}$

MC control + IS example

$\gamma = 0.9$



- $Q =$

w	x	y	z
-	4,3	2,1	-

- $C =$

w	x	y	z
-	0,0	0,0	-

- $\pi(s) =$

w	x	y	z
exit	←	←	exit

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s, a) \leftarrow$ arbitrary
  $C(s, a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

$\gamma = 0.9$

MC control + IS example

- $Q =$

w	x	y	z
-	4,3	2,1	-

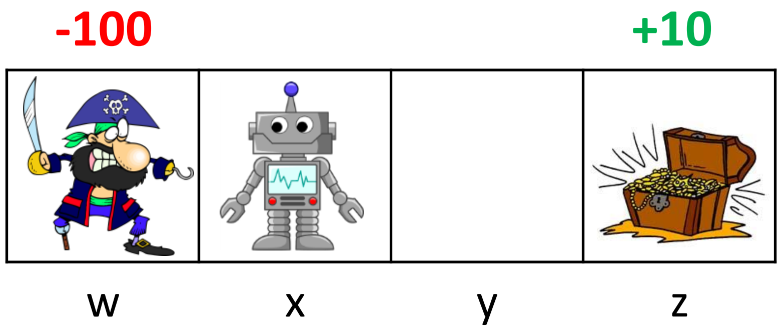
- $C =$

w	x	y	z
-	0,0	0,0	-

- $\pi(s) =$

w	x	y	z
exit	←	←	exit

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$



Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t,A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

- $Q =$

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

- $C =$

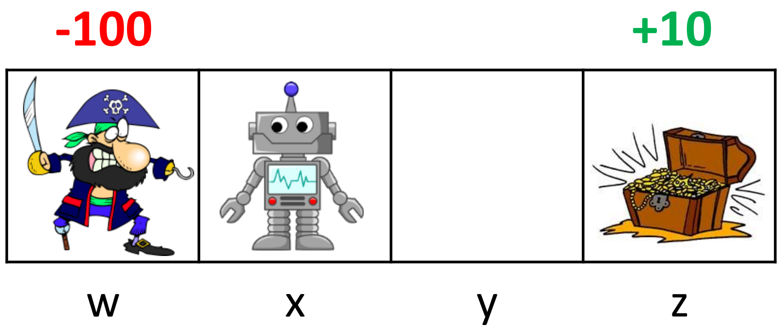
0	0,0	0,0	0
w	x	y	z

- $\pi(s) =$

exit	←	←	exit
w	x	y	z

- $b(s) = \forall s \; \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \leftarrow, 0, w, exit, -100$

$\gamma = 0.9$



Off-policy MC control, for estimating $\pi \approx \pi_*$

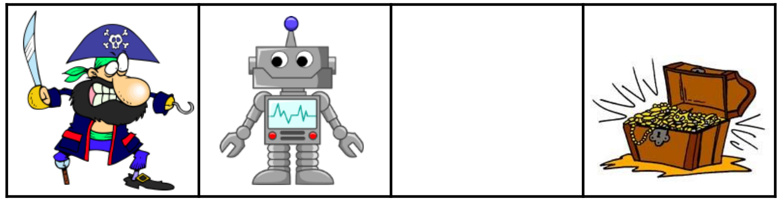
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$

-100 +10



w x y z

- $Q =$

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

- $C =$

0	0,0	0,0	0
w	x	y	z

- $\pi(s) =$

exit	←	←	exit
w	x	y	z

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \leftarrow, 0, w, exit, -100$

- $G = -100$
- $W = 1$
- $t = 1$

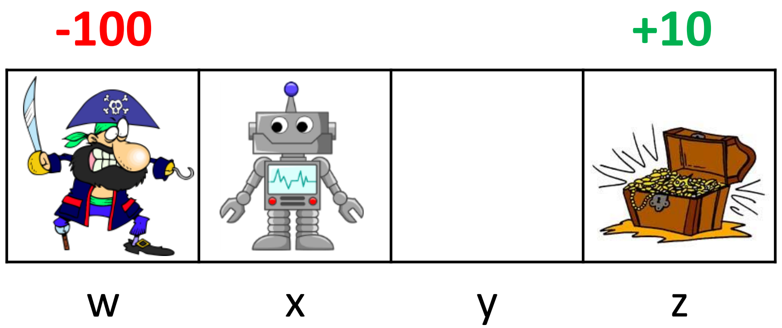
Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s, a) \leftarrow$ arbitrary
  $C(s, a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$



- $Q =$

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

- $C =$

w	x	y	z
1	0,0	0,0	0

- $\pi(s) =$

w	x	y	z
exit	←	←	exit

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \leftarrow, 0, w, exit, -100$

- $G = -100$
- $W = 1$
- $t = 1$

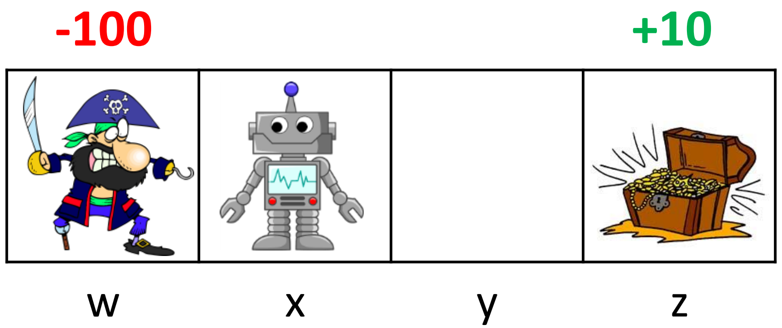
Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \mathrm{argmax}_a\, Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \mathrm{argmax}_a\, Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$



- $Q =$

-100	4,3	2,1	5
w	x	y	z

- $C =$

1	0,0	0,0	0
w	x	y	z

- $\pi(s) =$

exit	←	←	exit
w	x	y	z

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \leftarrow, 0, w, exit, -100$

- $G = -100$
- $W = 1$
- $t = 1$

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
$Q(s, a) \leftarrow$ arbitrary
$C(s, a) \leftarrow 0$
$\pi(s) \leftarrow \text{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
$b \leftarrow$ any soft policy
Generate an episode using b:
$S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
$G \leftarrow 0$
$W \leftarrow 1$
For $t = T-1, T-2, \ldots$ down to 0:
$G \leftarrow \gamma G + R_{t+1}$
$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
$\pi(S_t) \leftarrow \text{argmax}_a Q(S_t, a)$ (with ties broken consistently)
If $A_t \neq \pi(S_t)$ then exit For loop
$W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

- $Q =$

w	x	y	z
-100	4,3	2,1	5

- $C =$

w	x	y	z
1	0,0	0,0	0

- $\pi(s) =$

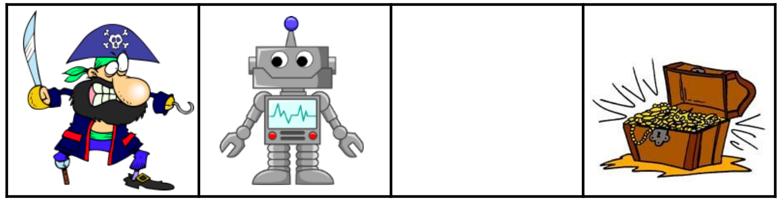
w	x	y	z
exit	←	←	exit

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \leftarrow, 0, w, exit, -100$

- $G = -100$
- $W = 1$
- $t = 1$

$\gamma = 0.9$

-100 +10



w x y z

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t,A_t)}[G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

- $Q =$

-100	4,3	2,1	5
w	x	y	z

- $C =$

1	0,0	0,0	0
w	x	y	z

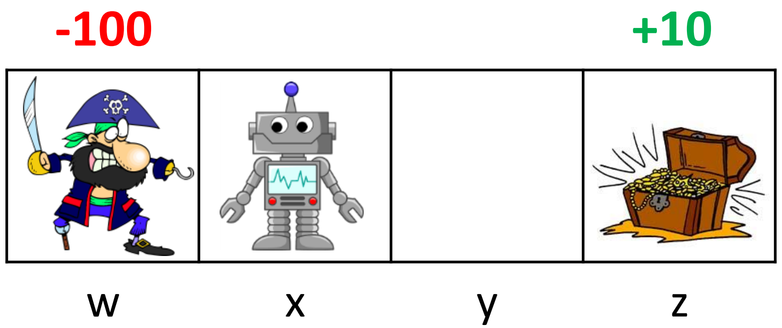
- $\pi(s) =$

exit	←	←	exit
w	x	y	z

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \leftarrow, 0, w, exit, -100$

- $G = -100$
- $W = 2$
- $t = 1$

$\gamma = 0.9$



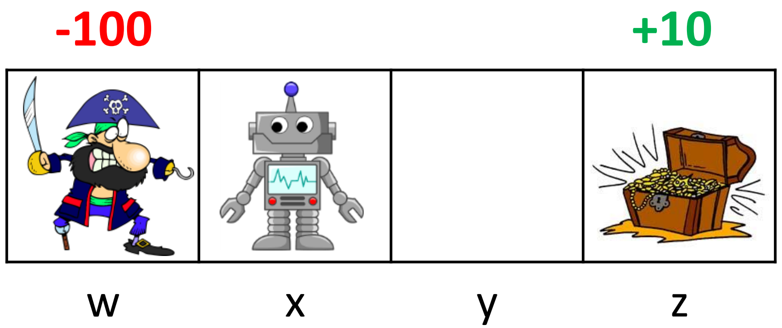
Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
$Q(s,a) \leftarrow$ arbitrary
$C(s,a) \leftarrow 0$
$\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
$b \leftarrow$ any soft policy
Generate an episode using b:
$S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
$G \leftarrow 0$
$W \leftarrow 1$
For $t = T-1, T-2, \ldots$ down to 0:
$G \leftarrow \gamma G + R_{t+1}$
$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
$\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
If $A_t \neq \pi(S_t)$ then exit For loop
$W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$



- $Q =$

w	x	y	z
-100	4,3	2,1	5

- $C =$

w	x	y	z
1	0,0	0,0	0

- $\pi(s) =$

w	x	y	z
exit	←	←	exit

- $b(s) = \forall s \; \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \leftarrow, 0, w, exit, -100$

- $G = -90$
- $W = 2$
- $t = 0$

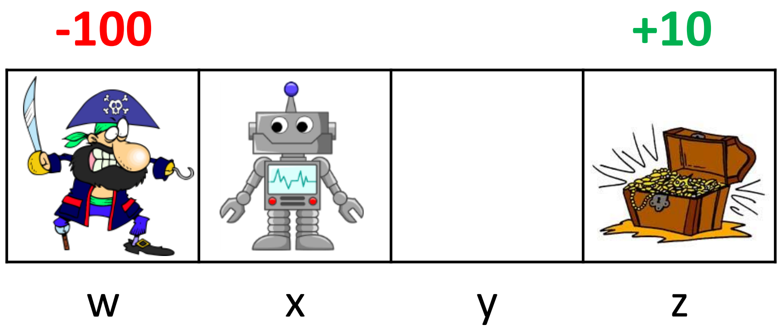
Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s, a) \leftarrow$ arbitrary
    $C(s, a) \leftarrow 0$
    $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
    $b \leftarrow$ any soft policy
    Generate an episode using b:
        $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
    $G \leftarrow 0$
    $W \leftarrow 1$
    For $t = T-1, T-2, \ldots$ down to 0:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit For loop
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$



- $Q =$

-100	4,3	2,1	5
w	x	y	z

- $C =$

1	2,0	0,0	0
w	x	y	z

- $\pi(s) =$

exit	←	←	exit
w	x	y	z

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \leftarrow, 0, w, exit, -100$

- $G = -90$
- $W = 2$
- $t = 0$

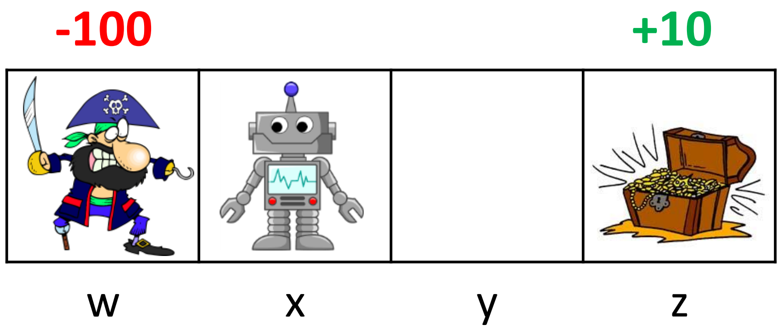
Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}\left[G - Q(S_t, A_t)\right]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$



- $Q =$

w	x	y	z
-100	-90,3	2,1	5

- $C =$

w	x	y	z
1	2,0	0,0	0

- $\pi(s) =$

w	x	y	z
exit	←	←	exit

- $b(s) = \forall s \,\{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \leftarrow, 0, w, exit, -100$

- $G = -90$
- $W = 2$
- $t = 0$

Off-policy MC control, for estimating $\pi \approx \pi_*$

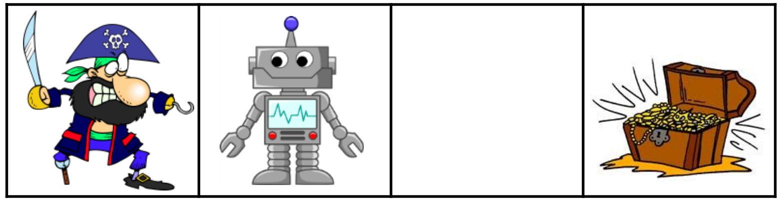
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$

-100 +10



w x y z

- $Q =$

w	x	y	z
-100	-90,3	2,1	5

- $C =$

w	x	y	z
1	2,0	0,0	0

- $\pi(s) =$

w	x	y	z
exit	→	←	exit

- $b(s) = \forall s \, \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \leftarrow, 0, w, exit, -100$

- $G = -90$
- $W = 2$
- $t = 0$

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
 $Q(s, a) \leftarrow$ arbitrary
 $C(s, a) \leftarrow 0$
 $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
 $b \leftarrow$ any soft policy
 Generate an episode using b:
  $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
 $G \leftarrow 0$
 $W \leftarrow 1$
 For $t = T-1, T-2, \ldots$ down to 0:
  $G \leftarrow \gamma G + R_{t+1}$
  $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
  $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$
  $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
  If $A_t \neq \pi(S_t)$ then exit For loop
  $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

- $Q =$

-100	-90,3	2,1	5
w	x	y	z

- $C =$

1	2,0	0,0	0
w	x	y	z

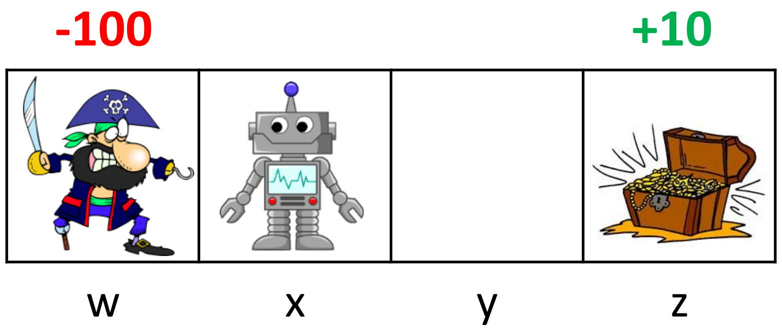
- $\pi(s) =$

exit	$\rightarrow$	$\leftarrow$	exit
w	x	y	z

- $b(s) = \forall s \ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \leftarrow, 0, w, exit, -100$

- $G = -90$
- $W = 2$
- $t = 0$

$\gamma = 0.9$



Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
$\quad Q(s,a) \leftarrow$ arbitrary
$\quad C(s,a) \leftarrow 0$
$\quad \pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
$\quad b \leftarrow$ any soft policy
$\quad$ Generate an episode using b:
$\quad\quad S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
$\quad G \leftarrow 0$
$\quad W \leftarrow 1$
$\quad$ For $t = T-1, T-2, \ldots$ down to 0:
$\quad\quad G \leftarrow \gamma G + R_{t+1}$
$\quad\quad C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
$\quad\quad Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
$\quad\quad \pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
$\quad\quad$ If $A_t \neq \pi(S_t)$ then exit For loop
$\quad\quad W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

- $Q =$

-100	-90,3	2,1	5
w	x	y	z

- $C =$

1	2,0	0,0	0
w	x	y	z

- $\pi(s) =$

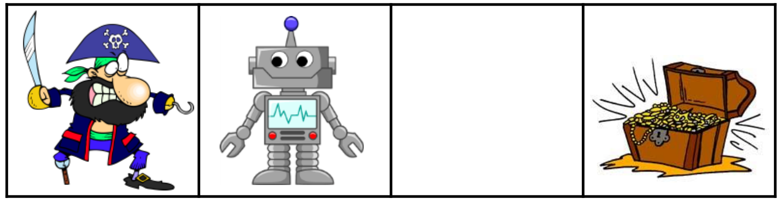
exit	→	←	exit
w	x	y	z

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = -90$
- $W = 2$
- $t = 0$

$\gamma = 0.9$

-100 +10



w x y z

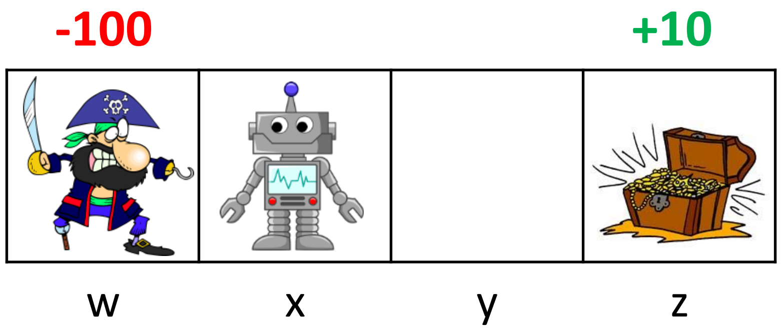
Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s, a) \leftarrow$ arbitrary
  $C(s, a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$



- $Q =$

w	x	y	z
-100	-90,3	2,1	5

- $C =$

w	x	y	z
1	2,0	0,0	0

- $\pi(s) =$

w	x	y	z
exit	→	←	exit

- $b(s) = \forall s \, \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 10$
- $W = 1$
- $t = 2$

Off-policy MC control, for estimating $\pi \approx \pi_*$

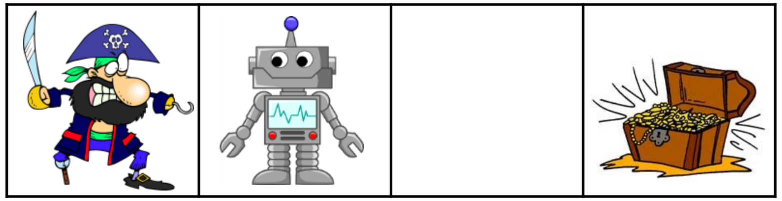
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s, a) \leftarrow$ arbitrary
  $C(s, a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$

-100 +10



w x y z

- $Q =$

-100	-90,3	2,1	5
w	x	y	z

- $C =$

1	2,0	0,0	1
w	x	y	z

- $\pi(s) =$

exit	→	←	exit
w	x	y	z

- $b(s) = \forall s \, \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 10$
- $W = 1$
- $t = 2$

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \leftarrow$ arbitrary
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
    $b \leftarrow$ any soft policy
    Generate an episode using b:
        $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
    $G \leftarrow 0$
    $W \leftarrow 1$
    For $t = T-1, T-2, \ldots$ down to 0:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit For loop
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

- $Q =$

-100	-90,3	2,1	10
w	x	y	z

- $C =$

1	2,0	0,0	1
w	x	y	z

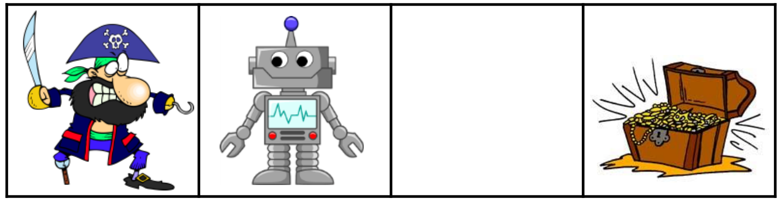
- $\pi(s) =$

exit	→	←	exit
w	x	y	z

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 10$
- $W = 1$
- $t = 2$

$\gamma = 0.9$



Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
 $Q(s,a) \leftarrow$ arbitrary
 $C(s,a) \leftarrow 0$
 $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
 $b \leftarrow$ any soft policy
 Generate an episode using b:
  $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
 $G \leftarrow 0$
 $W \leftarrow 1$
 For $t = T-1, T-2, \ldots$ down to 0:
  $G \leftarrow \gamma G + R_{t+1}$
  $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
  $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
  $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
  If $A_t \neq \pi(S_t)$ then exit For loop
  $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

- $Q =$

w	x	y	z
-100	-90,3	2,1	10

- $C =$

w	x	y	z
1	2,0	0,0	1

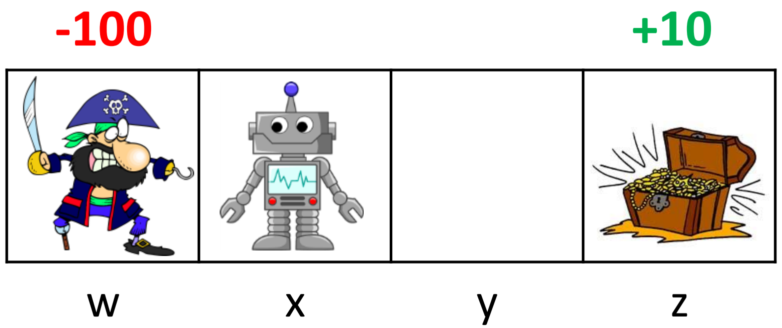
- $\pi(s) =$

w	x	y	z
exit	→	←	exit

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 10$
- $W = 1$
- $t = 2$

$\gamma = 0.9$



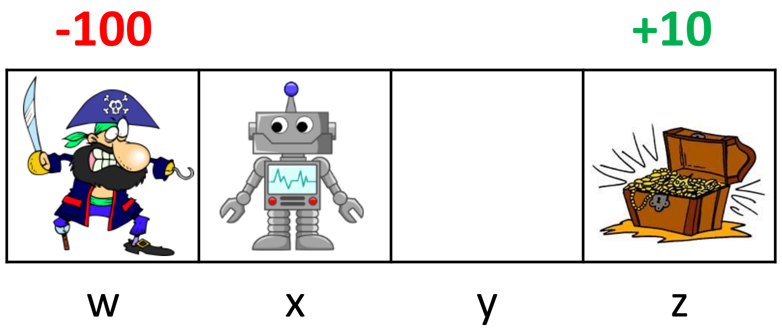
Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$



- $Q =$

w	x	y	z
-100	-90,3	2,1	10

- $C =$

w	x	y	z
1	2,0	0,0	1

- $\pi(s) =$

w	x	y	z
exit	→	←	exit

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 10$
- $W = 2$
- $t = 2$

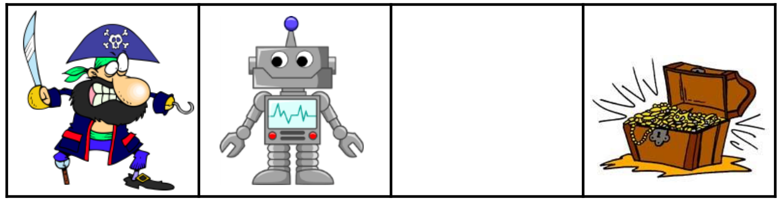
Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$



- $Q =$

w	x	y	z
-100	-90,3	2,1	10

- $C =$

w	x	y	z
1	2,0	0,0	1

- $\pi(s) =$

w	x	y	z
exit	→	←	exit

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 9$
- $W = 2$
- $t = 1$

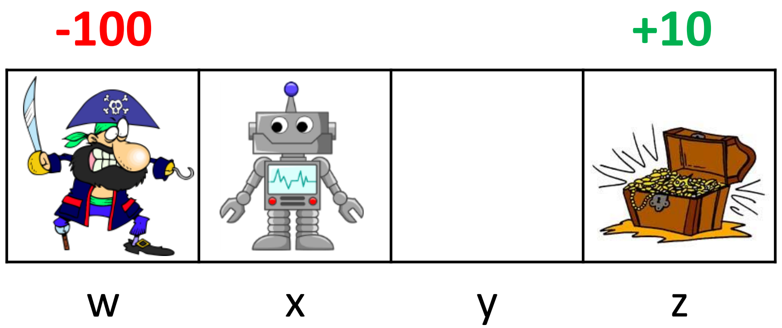
Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$



- $Q =$

w	x	y	z
-100	-90,3	2,1	10

- $C =$

w	x	y	z
1	2,0	0,2	1

- $\pi(s) =$

w	x	y	z
exit	$\rightarrow$	$\leftarrow$	exit

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 9$
- $W = 2$
- $t = 1$

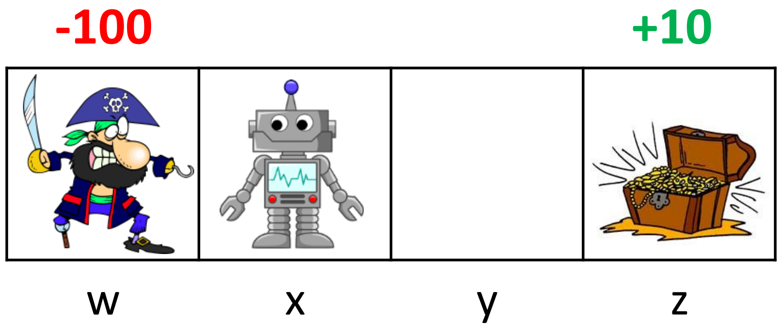
Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$



- $Q =$

-100	-90,3	2,9	10
w	x	y	z

- $C =$

1	2,0	0,2	1
w	x	y	z

- $\pi(s) =$

exit	→	←	exit
w	x	y	z

- $G = 9$
- $W = 2$
- $t = 1$

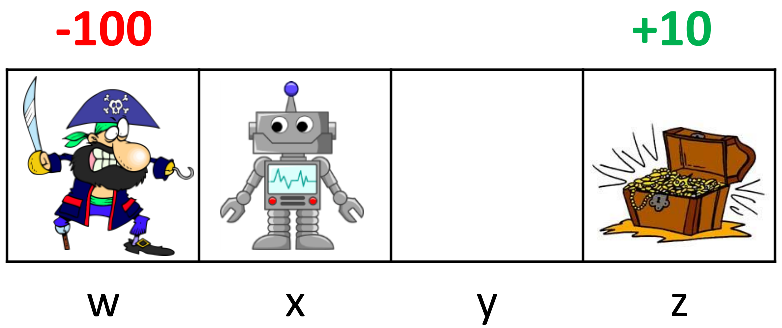
- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
 $Q(s,a) \leftarrow$ arbitrary
 $C(s,a) \leftarrow 0$
 $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
 $b \leftarrow$ any soft policy
 Generate an episode using b:
  $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
 $G \leftarrow 0$
 $W \leftarrow 1$
 For $t = T-1, T-2, \ldots$ down to 0:
  $G \leftarrow \gamma G + R_{t+1}$
  $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
  $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$
  $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
  If $A_t \neq \pi(S_t)$ then exit For loop
  $W \leftarrow W \frac{1}{b(A_t|S_t)}$

$\gamma = 0.9$



MC control + IS example

- $Q =$

w	x	y	z
-100	-90,3	2,9	10

- $C =$

w	x	y	z
1	2,0	0,2	1

- $\pi(s) =$

w	x	y	z
exit	$\rightarrow$	$\rightarrow$	exit

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 9$
- $W = 2$
- $t = 1$

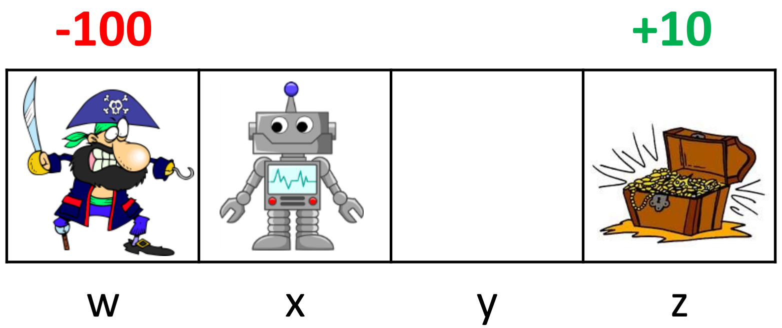
Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

$\gamma = 0.9$



- $Q =$

w	x	y	z
-100	-90,3	2,9	10

- $C =$

w	x	y	z
1	2,0	0,2	1

- $\pi(s) =$

w	x	y	z
exit	→	→	exit

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 9$
- $W = 4$
- $t = 1$

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \leftarrow$ arbitrary
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
    $b \leftarrow$ any soft policy
    Generate an episode using b:
        $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
    $G \leftarrow 0$
    $W \leftarrow 1$
    For $t = T-1, T-2, \ldots$ down to 0:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit For loop
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

- $Q =$

-100	-90,3	2,9	10
w	x	y	z

- $C =$

1	2,0	0,2	1
w	x	y	z

- $\pi(s) =$

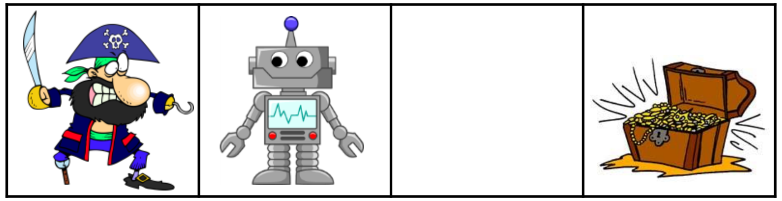
exit	$\rightarrow$	$\rightarrow$	exit
w	x	y	z

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 8.1$
- $W = 4$
- $t = 0$

$\gamma = 0.9$

-100 +10



w x y z

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \leftarrow$ arbitrary
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

- $Q =$

-100	-90,3	2,9	10
w	x	y	z

- $C =$

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

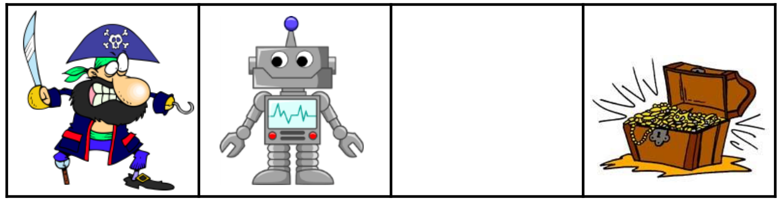
- $\pi(s) =$

exit	→	→	exit
w	x	y	z

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 8.1$
- $W = 4$
- $t = 0$

$\gamma = 0.9$



Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
 $Q(s, a) \leftarrow$ arbitrary
 $C(s, a) \leftarrow 0$
 $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
 $b \leftarrow$ any soft policy
 Generate an episode using b:
  $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
 $G \leftarrow 0$
 $W \leftarrow 1$
 For $t = T-1, T-2, \ldots$ down to 0:
  $G \leftarrow \gamma G + R_{t+1}$
  $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
  $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$
  $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
  If $A_t \neq \pi(S_t)$ then exit For loop
  $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

- $Q =$

w	x	y	z
-100	-90,8.1	2,9	10

- $C =$

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

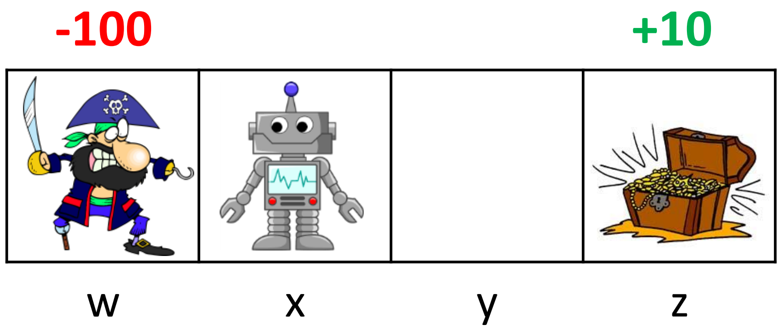
- $\pi(s) =$

w	x	y	z
exit	→	→	exit

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 8.1$
- $W = 4$
- $t = 0$

$\gamma = 0.9$



Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \leftarrow$ arbitrary
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t,a)$ (with ties broken consistently)

Repeat forever:
    $b \leftarrow$ any soft policy
    Generate an episode using b:
        $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
    $G \leftarrow 0$
    $W \leftarrow 1$
    For $t = T-1, T-2, \ldots$ down to 0:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t,A_t) \leftarrow C(S_t,A_t) + W$
        $Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \frac{W}{C(S_t,A_t)}[G - Q(S_t,A_t)]$
        $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t,a)$ (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit For loop
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

- $Q =$

-100	-90,8.1	2,9	10
w	x	y	z

- $C =$

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

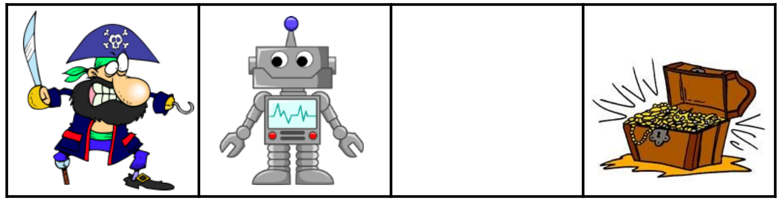
- $\pi(s) =$

exit	$\rightarrow$	$\rightarrow$	exit
w	x	y	z

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 8.1$
- $W = 8$
- $t = 0$

$\gamma = 0.9$



Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \leftarrow$ arbitrary
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
    $b \leftarrow$ any soft policy
    Generate an episode using b:
        $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
    $G \leftarrow 0$
    $W \leftarrow 1$
    For $t = T-1, T-2, \ldots$ down to 0:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit For loop
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$

MC control + IS example

- $Q =$

w	x	y	z
-100	-90,8.1	2,9	10

- $C =$

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

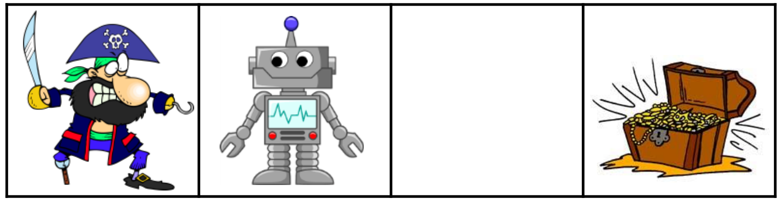
- $\pi(s) =$

w	x	y	z
exit	→	→	exit

- $b(s) = \forall s\ \{0.5 \leftarrow, 0.5 \rightarrow\}$
- $\tau = x, \rightarrow, 0, y, \rightarrow 0, z, exit, 10$

- $G = 8.1$
- $W = 8$
- $t = 0$

$\gamma = 0.9$



Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s, a) \leftarrow$ arbitrary
  $C(s, a) \leftarrow 0$
  $\pi(s) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
  $b \leftarrow$ any soft policy
  Generate an episode using b:
    $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  For $t = T-1, T-2, \ldots$ down to 0:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit For loop
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$

What did we learn?

- Online evaluation through the Monte-Carlo approach
  - Update V or Q estimations based on observed returns
- On policy Monte-Carlo control
  - Beware of local optimum. Must explore!
  - Consider using a soft policy $\pi(a|s)$
  - Assign soft policy values that mimic an ε-greedy strategy
  - On policy learning is not sample efficient!
- Off policy Monte-Carlo control
  - Define target policy (e.g., $\arg\max_a Q(S_t, a)$) that can be different than the behavior policy
  - Utilize weighted importance sampling to train a target policy
  - $v_\pi(s) = \frac{\sum_{\mathrm{m} \in M}[\rho_{t^m} G_t^m]}{\sum_{\mathrm{m} \in M} \rho_{t^m}}$

What next?

- **Lecture**: Temporal Difference Learning
- **Assignments**:
  - Monte-Carlo Control
  - Monte-Carlo Control with Importance Sampling
  - Due by Monday, ?, EOD
- **Quiz (on Canvas)**:
  - Monte-Carlo Control
  - By Sunday, ?, EOD
- **Project**:
  - Converge on your project’s topic and scope