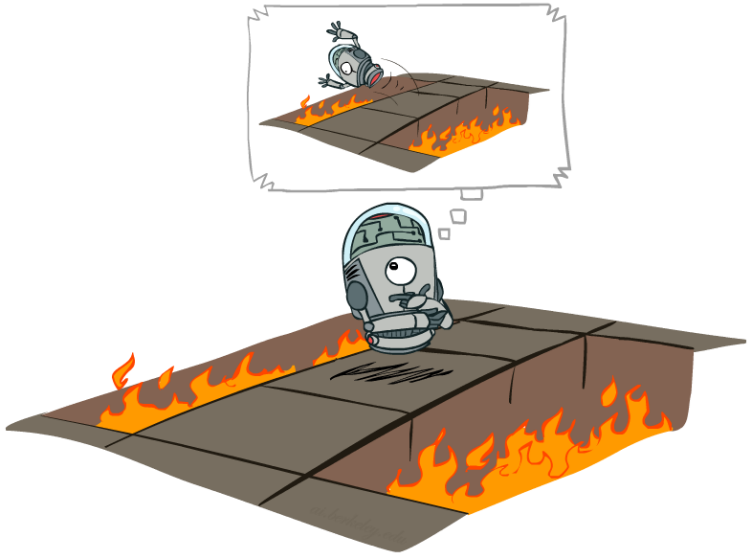
(Aside) Offline vs. Online (RL)



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

1. First-Visit Monte Carlo Method
2. Every-Visit Monte Carlo Method

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

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 $Returns(s)$

$V(s) \leftarrow$ average($Returns(s)$)

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

5.1 Monte Carlo Prediction

We begin by considering Monte Carlo methods for learning the state-value function for a given policy. Recall that the value of a state is the expected return—expected cumulative future discounted reward—starting from that state. An obvious way to estimate it from experience, then, is simply to average the returns observed after visits to that state. As more returns are observed, the average should converge to the expected value. This idea underlies all Monte Carlo methods.

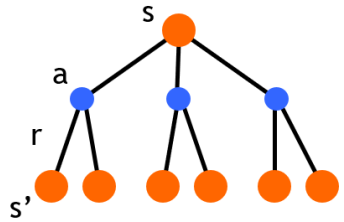
In particular, suppose we wish to estimate $v_\pi(s)$, the value of a state s under policy π, given a set of episodes obtained by following π and passing through s. Each occurrence of state s in an episode is called a *visit* to s. Of course, s may be visited multiple times in the same episode; let us call the first time it is visited in an episode the *first visit* to s. The *first-visit MC method* estimates $v_\pi(s)$ as the average of the returns following first visits to s, whereas the *every-visit MC method* averages the returns following all visits to s. These two Monte Carlo (MC) methods are very similar but have slightly different theoretical properties. First-visit MC has been most widely studied, dating back to the 1940s, and is the one we focus on in this chapter. Every-visit MC extends more naturally to function approximation and eligibility traces, as discussed in Chapters 9 and 12. First-visit MC is shown in procedural form in the box.

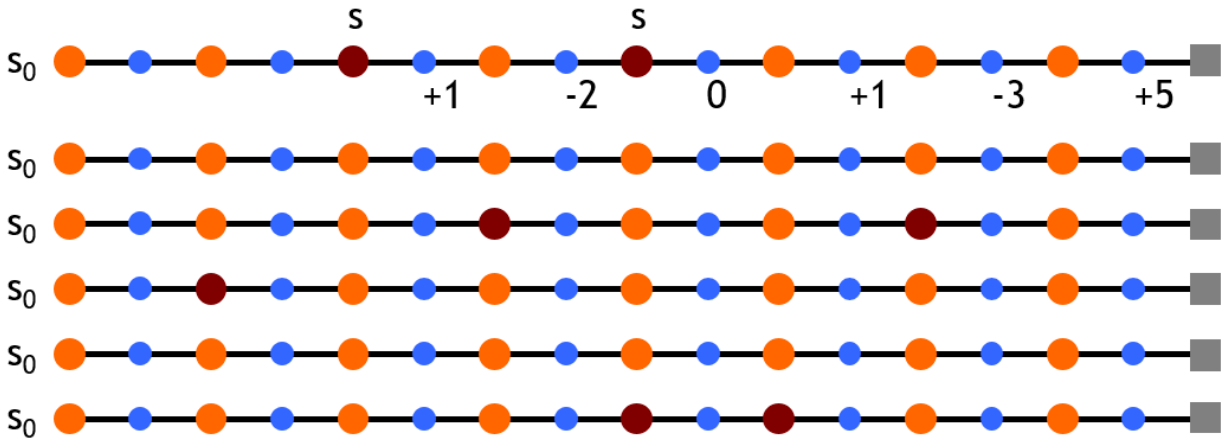
First-visit MC prediction, for estimating $V \approx v_\pi$

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:
    Generate an episode using π
    For each state s appearing in the episode:
        $G \leftarrow$ the return that follows the first occurrence of s
        Append G to $Returns(s)$
        $V(s) \leftarrow$ average($Returns(s)$)

Ex-1:First-visit Monte-Carlo Policy Evaluation [estimate $V\pi(s)$]





$R_1(s) = +2$

$R_2(s) = +1$

$R_3(s) = -5$

$R_4(s) = +4$

$$V^{\pi}(s) \approx (2 + 1 - 5 + 4)/4 = 0.5$$

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

Ex-2:First-visit Monte-Carlo Policy Evaluation [estimate Vπ(s)]

E1: A+3->A+2->B-4->A+4->B-3->Terminate
E2: B-2->A+3->B-3->Terminate

Given Episodes

1. **Episode 1 (E1)**: A 3 A 2 B -4 A 4 B -3 T
2. **Episode 2 (E2)**: B -2 A 3 B -3 T

Episode 1: A 3 A 2 B -4 A 4 B -3 T

- **First Visit of** A:
  - **Return** G: From the first visit to A (at time step 0):

    $G = 3(\text{reward at } A) + 2(\text{reward at } A) + (-4)(\text{reward at } B) + 4(\text{reward at } A) + (-3)$(reward at B)

  - Append G to $Returns(A)$: $Returns(A) = [2]$
- **First Visit of** B:
  - **Return** G: From the first visit to B (at time step 2):

    $G = -4(\text{reward at } B) + 4(\text{reward at } A) + (-3)(\text{reward at } B) = -4 + 4 - 3 = -3$

  - Append G to $Returns(B)$: $Returns(B) = [-3]$

Ex-2:First-visit Monte-Carlo Policy Evaluation [estimate Vπ(s)]

E1: A+3->A+2->B-4->A+4->B-3->Terminate
E2: B-2->A+3->B-3->Terminate

Given Episodes

1. **Episode 1 (E1)**: A 3 A 2 B -4 A 4 B -3 T
2. **Episode 2 (E2)**: B -2 A 3 B -3 T

Episode 2: B -2 A 3 B -3 T

- **First Visit of B**:
  - **Return G**: From the first visit to B (at time step 0):

    $$G = -2(\text{reward at } B) + 3(\text{reward at } A) + (-3)(\text{reward at } B) = -2 + 3 - 3 = -2$$

  - Append G to $Returns(B)$: $Returns(B) = [-3, -2]$
- **First Visit of A**:
  - **Return G**: From the first visit to A (at time step 1):

    $$G = 3(\text{reward at } A) + (-3)(\text{reward at } B) = 3 - 3 = 0$$

  - Append G to $Returns(A)$: $Returns(A) = [2, 0]$

Ex-2:First-visit Monte-Carlo Policy Evaluation [estimate Vπ(s)]

E1: A+3->A+2->B-4->A+4->B-3->Terminate

E2: B-2->A+3->B-3->Terminate

Given Episodes

1. Episode 1 (E1): A 3 A 2 B -4 A 4 B -3 T
2. Episode 2 (E2): B -2 A 3 B -3 T

Calculate Value Estimates

1. For A:
   - Returns: $Returns(A) = [2, 0]$
   - Average Value $V(A)$:

$$V(A) = \frac{2+0}{2} = \frac{2}{2} = 1$$

2. For B:
   - Returns: $Returns(B) = [-3, -2]$
   - Average Value $V(B)$:

$$V(B) = \frac{-3+(-2)}{2} = \frac{-5}{2} = -2.5$$

Summary

- Estimated Value of State A: 1
- Estimated Value of State B: -2.5

Ex-3:First-visit Monte-Carlo Policy Evaluation [estimate Vπ(s)]

Exercise 3: Assume you have data in the form of just the following 5 complete episodes. Non-terminal States are labeled A and B, the numbers in the episodes denote Rewards, and all states end in a terminal state T.

- A 2 A 6 B 1 B 0 T
- A 3 B 2 A 4 B 2 B 0 T
- A 0 B 2 A 4 B 4 B 2 B 0 T
- B 8 B 0 T

Estimate the value of states of A & B. Show the steps

Given Episodes

1. $A\ 2\ A\ 6\ B\ 1\ B\ 0\ T$
2. $A\ 3\ B\ 2\ A\ 4\ B\ 2\ B\ 0\ T$
3. $A\ 0\ B\ 2\ A\ 4\ B\ 4\ B\ 2\ B\ 0\ T$
4. $B\ 8\ B\ 0\ T$

Ex-3:First-visit Monte-Carlo Policy Evaluation [estimate Vπ(s)]

Given Episodes

1. $A\ 2\ A\ 6\ B\ 1\ B\ 0\ T$
2. $A\ 3\ B\ 2\ A\ 4\ B\ 2\ B\ 0\ T$
3. $A\ 0\ B\ 2\ A\ 4\ B\ 4\ B\ 2\ B\ 0\ T$
4. $B\ 8\ B\ 0\ T$

Episode 1: $A\ 2\ A\ 6\ B\ 1\ B\ 0\ T$

- **First Visit of** A:
  - **Return** G: From the first occurrence of A:

    $$G = 2(\text{reward at } A) + 6(\text{reward at } A) + 1(\text{reward at } B) + 0(\text{reward at } B) = 9$$

  - Append G to $Returns(A)$: $Returns(A) = [9]$
- **First Visit of** B:
  - **Return** G: From the first occurrence of B:

    $$G = 1(\text{reward at } B) + 0(\text{reward at } B) = 1$$

  - Append G to $Returns(B)$: $Returns(B) = [1]$

Ex-3:First-visit Monte-Carlo Policy Evaluation [estimate Vπ(s)]

Given Episodes

1. $A\ 2\ A\ 6\ B\ 1\ B\ 0\ T$
2. $A\ 3\ B\ 2\ A\ 4\ B\ 2\ B\ 0\ T$
3. $A\ 0\ B\ 2\ A\ 4\ B\ 4\ B\ 2\ B\ 0\ T$
4. $B\ 8\ B\ 0\ T$

Episode 2: $A\ 3\ B\ 2\ A\ 4\ B\ 2\ B\ 0\ T$

- **First Visit of** A:
  - **Return** G: From the first occurrence of A:

    $$G = 3(\text{reward at } A) + 2(\text{reward at } B) + 4(\text{reward at } A) + 2(\text{reward at } B) = 11$$

  - Append G to $Returns(A)$: $Returns(A) = [9, 11]$
- **First Visit of** B:
  - **Return** G: From the first occurrence of B:

    $$G = 2(\text{reward at } B) + 4(\text{reward at } A) + 2(\text{reward at } B) + 0(\text{reward at } B) = 8$$

  - Append G to $Returns(B)$: $Returns(B) = [1, 8]$

Ex-3:First-visit Monte-Carlo Policy Evaluation [estimate Vπ(s)]

Given Episodes

1. $A\ 2\ A\ 6\ B\ 1\ B\ 0\ T$
2. $A\ 3\ B\ 2\ A\ 4\ B\ 2\ B\ 0\ T$
3. $A\ 0\ B\ 2\ A\ 4\ B\ 4\ B\ 2\ B\ 0\ T$
4. $B\ 8\ B\ 0\ T$

Episode 3: $A\ 0\ B\ 2\ A\ 4\ B\ 4\ B\ 2\ B\ 0\ T$

- **First Visit of** A:
  - **Return** G: From the first occurrence of A:

    $$G = 0(\text{reward at } A) + 2(\text{reward at } B) + 4(\text{reward at } A) + 4(\text{reward at } B) + 2(\text{reward at } B) + 0(\text{reward at } B) = 12$$

  - Append G to $Returns(A)$: $Returns(A) = [9, 11, 12]$
- **First Visit of** B:
  - **Return** G: From the first occurrence of B:

    $$G = 2(\text{reward at } B) + 4(\text{reward at } A) + 4(\text{reward at } B) + 2(\text{reward at } B) + 0(\text{reward at } B) = 12$$

  - Append G to $Returns(B)$: $Returns(B) = [1, 8, 12]$

Ex-3:First-visit Monte-Carlo Policy Evaluation [estimate Vπ(s)]

Given Episodes

1. $A\ 2\ A\ 6\ B\ 1\ B\ 0\ T$
2. $A\ 3\ B\ 2\ A\ 4\ B\ 2\ B\ 0\ T$
3. $A\ 0\ B\ 2\ A\ 4\ B\ 4\ B\ 2\ B\ 0\ T$
4. $B\ 8\ B\ 0\ T$

Episode 4: $B\ 8\ B\ 0\ T$

- **First Visit of** B:
  - **Return** G: From the first occurrence of B:

$$G = 8(\text{reward at } B) + 0(\text{reward at } B) = 8$$

  - Append G to $Returns(B)$: $Returns(B) = [1, 8, 12, 8]$

Ex-3:First-visit Monte-Carlo Policy Evaluation [estimate Vπ(s)]

Given Episodes

1. $A\,2\,A\,6\,B\,1\,B\,0\,T$
2. $A\,3\,B\,2\,A\,4\,B\,2\,B\,0\,T$
3. $A\,0\,B\,2\,A\,4\,B\,4\,B\,2\,B\,0\,T$
4. $B\,8\,B\,0\,T$

Calculate Value Estimates

1. For A:
   - Returns: $Returns(A) = [9, 11, 12]$
   - Average Value $V(A)$:

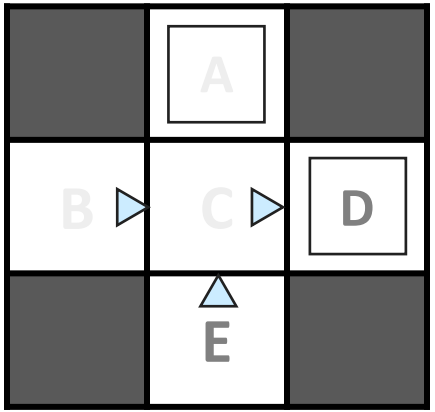
$$V(A) = \frac{9 + 11 + 12}{3} = \frac{32}{3} \approx 10.67$$

2. For B:
   - Returns: $Returns(B) = [1, 8, 12, 8]$
   - Average Value $V(B)$:

$$V(B) = \frac{1 + 8 + 12 + 8}{4} = \frac{29}{4} = 7.25$$

Ex-4: First-visit Monte-Carlo Policy Evaluation [estimate Vπ(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

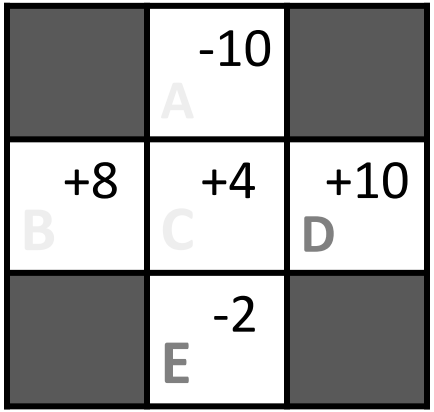
Episode 3

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

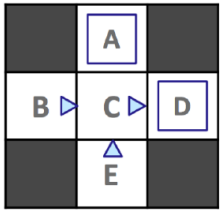
Episode 4

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

Output Values



Ex-4: First-visit Monte-Carlo Policy Evaluation [estimate Vπ(s)]



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

Episode 2
B, east, C, -1
C, east, D, -1
D, exit, x, +10

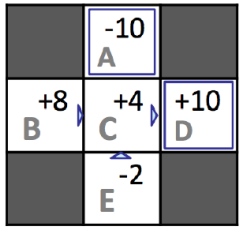
Episode 3
E, north, C, -1
C, east, D, -1
D, exit, x, +10

Episode 4
E, north, C, -1
C, east, A, -1
A, exit, x, -10

Walking through the first episode, we can see that from state D to termination we acquired a total reward of 10, from state C we acquired a total reward of $(-1) + 10 = 9$, and from state B we acquired a total reward of $(-1) + (-1) + 10 = 8$. Completing this process yields the total reward across episodes for each state and the resulting estimated values as follows:

s	Total Reward	Times Visited	$V^{\pi}(s)$
A	-10	1	-10
B	16	2	8
C	16	4	4
D	30	3	10
E	-4	2	-2

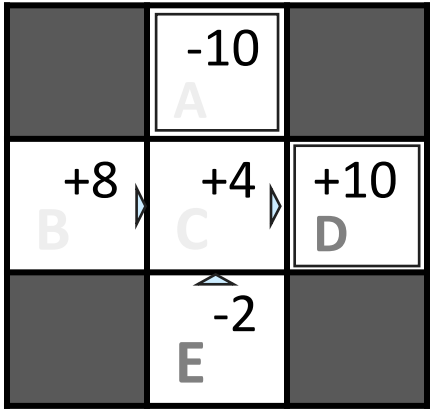
Though direct evaluation eventually learns state values for each state, it's often unnecessarily slow to converge because it wastes information about transitions between states.



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?

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

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

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Thank you