

SMU: Lecture 2

(Model-Free Policy Evaluation in RL)

Monday, February 14, 2022

(Heavily inspired by the Stanford RL Course of Prof. Emma Brunskill, but all potential errors are mine.)

Plan for Today

- Policy evaluation when we do not know the model (neither the state-transition probabilities, nor the reward functions).
- Two kinds of methods today (there are more out there):
 - Monte-Carlo Policy Evaluation
 - Temporal-Difference Learning
- *Note: The promised material about finite-horizon problems will not fit in today's lecture (but I will make a video about it).*

Part 0: Reminder from Last Lecture

Markov Reward Process

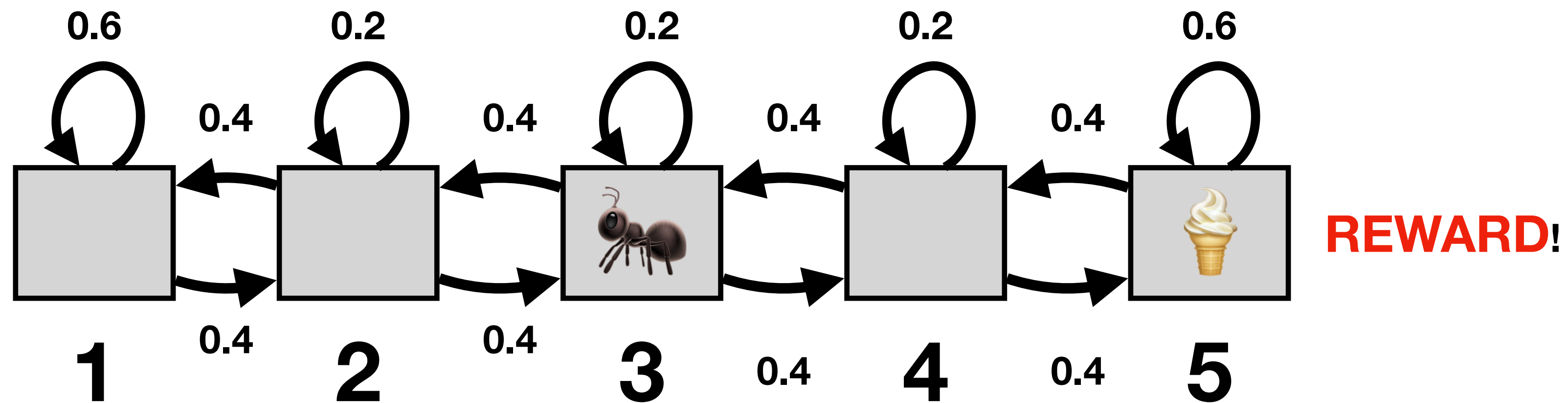
Markov reward process = Markov process + Reward

Formally, MRP is given by:

- A set of states S .
- A transition model $P[X_{t+1} = s' | X_t = s]$, which we also denote by $P(s' | s)$.
- A reward function $R(s) = \mathbb{E}[R_t | X_t = s]$, which is the expected reward the agent receives in state s , ($s \in S$).
- A discount factor $\gamma \in [0; 1]$.

Markov Reward Process

Markov reward process = Markov process + Reward



For example:

$$R(s) = \begin{cases} 0, & s = 1 \\ 0, & s = 2 \\ 0, & s = 3 \\ 0, & s = 4 \\ 10, & s = 5 \end{cases}$$

← We expect that each time we visit s_5 , there will be ice cream (i.e. we are not running out of it).

Return from an Episode

- **Horizon:**
 - Number of time steps in an episode (which can also be infinite). **We will first assume infinite horizons** (they are easier because they will lead to stationary, i.e. time-independent, policies!).
- **Return G_t :**
 - **Given:** An episode $s_1, s_2, s_3, s_4, \dots, s_H$.
 - **Compute:** Return $g_t =$ discounted sum of rewards from time t .
 - **As a formula:**

$$g_t = R(s_t) + R(s_{t+1}) \cdot \gamma + R(s_{t+2}) \cdot \gamma^2 + \dots = R(s_t) + \sum_{i=1} R(s_{t+i}) \cdot \gamma^i$$

Return (Random Variable)

- What we had on the previous slide was return from one specific sampled episode.
- Next we define **return** of a Markov reward process as a random variable (it is important to understand the distinction between the two):

$$G_t = R(X_t) + \gamma \cdot R(X_{t+1}) + \gamma^2 \cdot R(X_{t+2}) + \dots = \sum_{i=0}^{\infty} R(X_{t+i}) \cdot \gamma^i$$

Markov Decision Process


- **Markov decision process = Markov reward process + Actions**
- **An MDP is given by:**

- A set of states S .
- A set of actions A .

- A transition model $P[X_{t+1} = s' | X_t = s, A_t = a] = \underbrace{P(s' | s, a)}_{\text{notation}}$

- A reward $R(s, a) = \mathbb{E}[R_t | X_t = s, A_t = a]$, i.e. the expected reward that the agent receives when performing action a in state s .
- Discount factor γ .

Policy

- Policy determines which action to take in each state s .
- It can be either deterministic or random — that is also why policy will not simply be a function from states to actions.
- **We define policy:** $\pi(a | s) = P(A_t = a | X_t = s)$.
- **Example** (policy for our ant 

9

MDP+Policy = MRP

- When we specify a policy for a given MDP, we are effectively turning the MDP into a corresponding MRP.
- **Formally:**
 - Given an MDP (A, S, P, R, γ) , we turn it into an MRP $(S, P^\pi, R^\pi, \gamma)$ where

$$P^\pi(s' | s) = \sum_{a \in A} \pi(a | s) \cdot P(s' | s, a) *$$

$$R^\pi(s) = \sum_{a \in A} \pi(a | s) \cdot R(s, a)$$

* In the more verbose notation: $P^\pi[X_{t+1} = s' | X_t = s] = \sum_{a \in A} \pi(a | s) \cdot P[X_{t+1} = s' | A_t = a, X_t = s]$.

State Value Function of MDP

General case:

$$V^\pi(s) = \sum_{a \in A} \pi(a, s) \cdot \left[R(s, a) + \gamma \cdot \sum_{s' \in S} P(s' | s, a) \cdot V^\pi(s') \right]$$

Version for deterministic policy:

$$V^\pi(s) = R(s, \pi(s)) + \gamma \cdot \sum_{s' \in S} P(s' | s, \pi(s)) \cdot V^\pi(s')$$

Part 1: Problem Statement

Problem: Model-Free Policy Evaluation

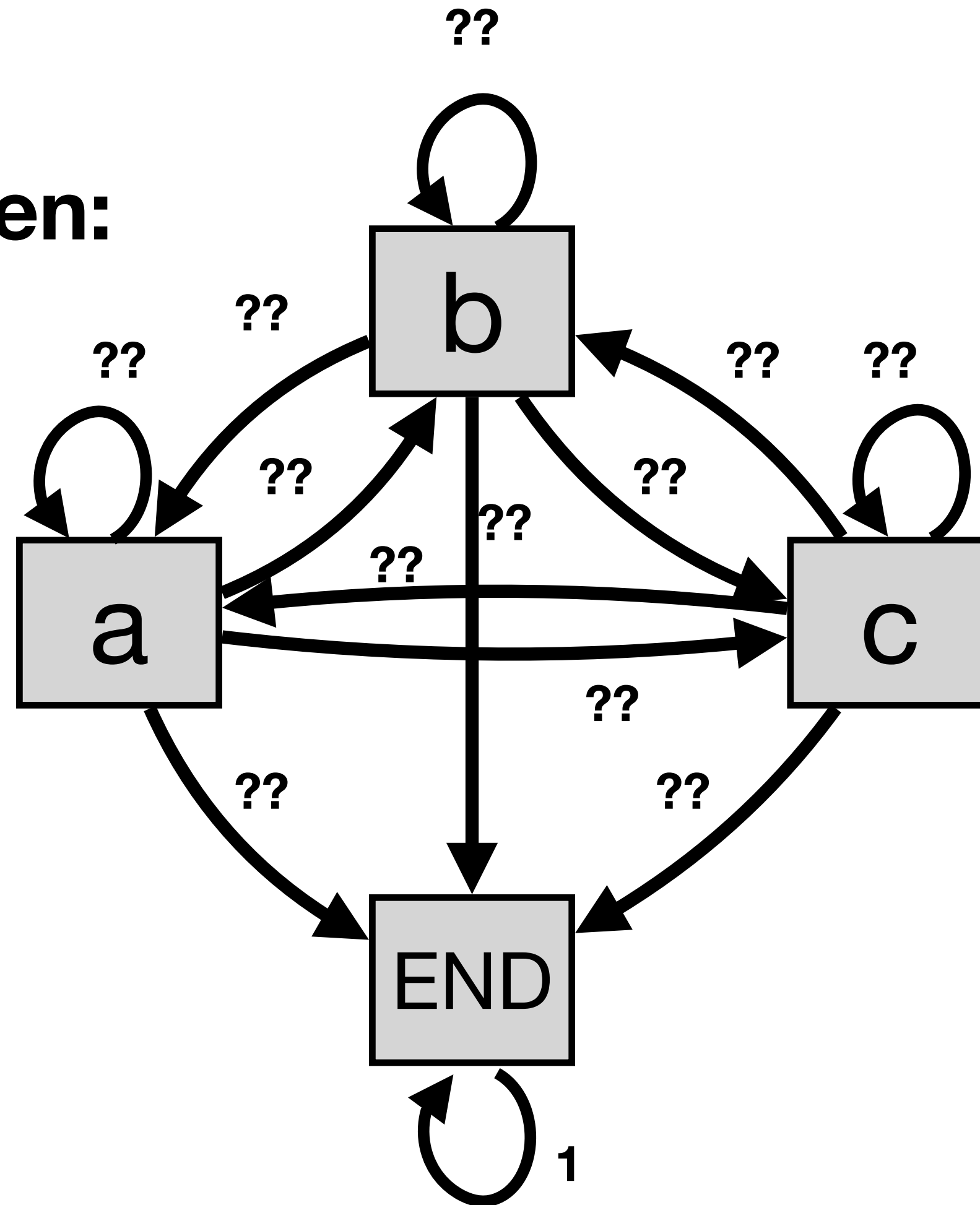
- Given a policy and an MDP with unknown parameters (or generally an environment with which we can interact), **estimate the value function.**

Example

Agent: 

Rewards??

States are given:



Actions are given:

$$A = \{l, r\}$$



Policy is given, e.g.:

$$\pi(l | a) = 0.2, \pi(r | a) = 0.8,$$
$$\pi(l | b) = 0.3, \pi(r | b) = 0.7,$$

...

An Assumption

- **Assumption:** In what follows we will assume that our MDP has terminal states and that the probability of infinitely long runs is zero.
- **Terminal states:** Once the system gets into a terminal state, it stays in it. The reward in the terminal state is always 0.
- **Why do we do this?** This assumption will allow us to use the formalism for infinite-horizon problems (which is mathematically simpler), yet it will simplify some things we will be talking about today.

Part 2: Statistical Properties of Estimators

(An informal recap of what you already know from statistics)

Estimators (Statistics)

- Suppose that we want to estimate a parameter θ . An **estimator of θ** is a function $\hat{\theta}$ that maps samples to estimates of the parameter θ .

Estimators as Random Variables

- **Example:** Let us have a normal distribution with mean μ and standard deviation σ . Denote by $\mathbf{X} = (X_1, X_2, \dots, X_N)$ an independent sample from this distribution. Then the sample mean $\hat{\mu}(\mathbf{X}) = \frac{1}{N} \sum_{i=1}^N X_i$ is an estimator for the population mean μ .
- Note that, in this example, $\hat{\mu}(\mathbf{X})$ is a **random variable**.

Bias

Bias of an estimator $\hat{\theta}$ is defined as: $\text{BIAS}_{\theta}(\hat{\theta}) = \mathbb{E}_{\theta}[\hat{\theta}(\mathbf{X})] - \theta$.

If $\text{BIAS}_{\theta}(\hat{\theta}) = 0$ then we say that $\hat{\theta}$ is an unbiased estimator.

Example: $\frac{1}{N} \sum_{k=1}^N X_k$ is an unbiased estimator of population mean. Why?

Because we have $\mathbb{E} \left[\frac{1}{N} \sum_{k=1}^N X_k \right] = \frac{1}{N} \sum_{k=1}^N \mathbb{E} [X_k] = \frac{1}{N} \cdot N \cdot \mathbb{E} [X_k] = \mu$.

Mean Squared Error

Mean squared error of an estimator $\hat{\theta}$ is defined as: $MSE_{\theta}(\hat{\theta}) = \mathbb{E}_{\theta}[(\hat{\theta}(\mathbf{X}) - \theta)^2]$.

It holds $MSE_{\theta}(\hat{\theta}(\mathbf{X})) = \text{Var}_{\theta}(\hat{\theta}(\mathbf{X})) + \text{BIAS}(\hat{\theta}(\mathbf{X}))^2$.

Consistency

Let $\mathbf{X} = (X_1, \dots, X_N)$ be an independent sample, used to estimate θ .

A sequence of estimators $\hat{\theta}_N(\mathbf{X})$ is said to be consistent if for every $\varepsilon > 0$ it holds: $\lim_{N \rightarrow \infty} P[|\hat{\theta}_N(\mathbf{X}) - \theta| < \varepsilon] = 1$.

Why It Matters

- Estimators that we are going to study in this lecture can be analyzed in the same framework. After all, they are just statistical estimators.

Part 3: Monte-Carlo Policy Evaluation

Monte-Carlo Policy Evaluation (1/5)

Recall the definition of G_t , the return at time t (*we have not shown it explicitly for MDPs last time*):

$$G_t^\pi = R(X_t, A_t) + \gamma \cdot R(X_{t+1}, A_{t+1}) + \gamma^2 \cdot R(X_{t+2}, A_{t+2}) + \dots = \sum_{i=0}^{\infty} R(X_{t+i}, A_{t+i}) \cdot \gamma^i$$

(for simplicity, we assume that the reward when $R(a,s)$ is deterministic)

where X_i 's and A_i 's are random variables — X_i is the state at time t and A_i is the action at time i . We suppose that these random variables are from an MDP with a policy π (which together define the distribution of these random variables).

Monte-Carlo Policy Evaluation (2/5)

The state value function $V^\pi(s)$ is:

$$V^\pi(s) = \mathbb{E}[G_t^\pi | X_t = s].$$

We were computing $V^\pi(s)$ by solving the Bellman equation (directly or iteratively):

$$V^\pi(s) = \sum_{a \in A} \pi(a, s) \cdot \left[R(s, a) + \gamma \cdot \sum_{s' \in S} P(s' | s, a) \cdot V^\pi(s') \right].$$

But there is also another way to approximate $V^\pi(s)$. *

**This method will not be very efficient for MDPs but bear with me... we are getting somewhere)*

Monte-Carlo Policy Evaluation (3/5)

An **episode** sampled from an MDP under a policy π is a sequence of states, actions and rewards which ends in a terminal state:

$$s_1, a_1, r_1, s_2, a_2, r_2, s_3, a_3, r_3, \dots, s_T$$

where s_i is the state at time i , a_i is the action taken at time i and r_i is the corresponding reward obtained at time i .

The return at time t for a concrete episode $s_1, a_1, s_2, a_2, \dots$

$$g_t = R(x_t, a_t) + \gamma \cdot R(x_{t+1}, a_{t+1}) + \gamma^2 \cdot R(x_{t+2}, a_{t+2}) + \dots = \sum_{i=0}^{\infty} R(x_{t+i}, a_{t+i}) \cdot \gamma^i$$

(for simplicity, we assume that the reward when $R(a,s)$ is deterministic but it can also be random... nothing will change regarding the algorithms, just the analysis would change little bit, but mostly just regarding notation)

We can have bounds ∞ , just remember that all rewards after T are 0.

where x_i 's are states from S and a_i 's are actions from A .

Monte-Carlo Policy Evaluation (4/5)

We will now try to approximate $V^\pi(s)$ directly using $V^\pi(s) = \mathbb{E}[G_t^\pi | X_t = s]$ using sampled episodes. *After all, expectation can be approximated by an average of sampled values.*

We will sample finite episodes (after all we can't sample infinitely long episodes in practice). *This also means that MC policy estimation can only be used for episodic RL problems.*

Monte-Carlo Policy Evaluation (5/5)

Why the problem is not straightforward: *If we only wanted to estimate $\mathbb{E}[G_t]$, that would be easy, but we want to estimate $\mathbb{E}[G_t | X_t = s]$ that is we need to condition... but we cannot condition arbitrarily... we can only observe episodes sampled under the given policy... **so we will need to “wait” for s to occur.***

We will see two different MC algorithms to do that: First-Visit MC Estimation and Every-Visit MC Estimation.

First-Visit Monte-Carlo Evaluation

Initialize: $G(s) = 0$, $N(s) = 0$ for all $s \in S$.

For $i = 1, \dots, N$:

Sample episode $e_i := s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \dots, s_{i,T_i}$.

For each time step $1 \leq t \leq T_i$:

If t is the first occurrence of state s in the episode e_i

$$g_{i,t} := r_{i,t} + \gamma \cdot r_{i,t+1} + \gamma^2 \cdot r_{i,t+2} + \dots + \gamma^{T_i-t} \cdot r_{i,T_i}$$

$$N(s) := N(s) + 1 \text{ /* Increment total visits counter */}$$

$$G(s) := G(s) + g_{i,1} \text{ /* Increment total return counter */}$$

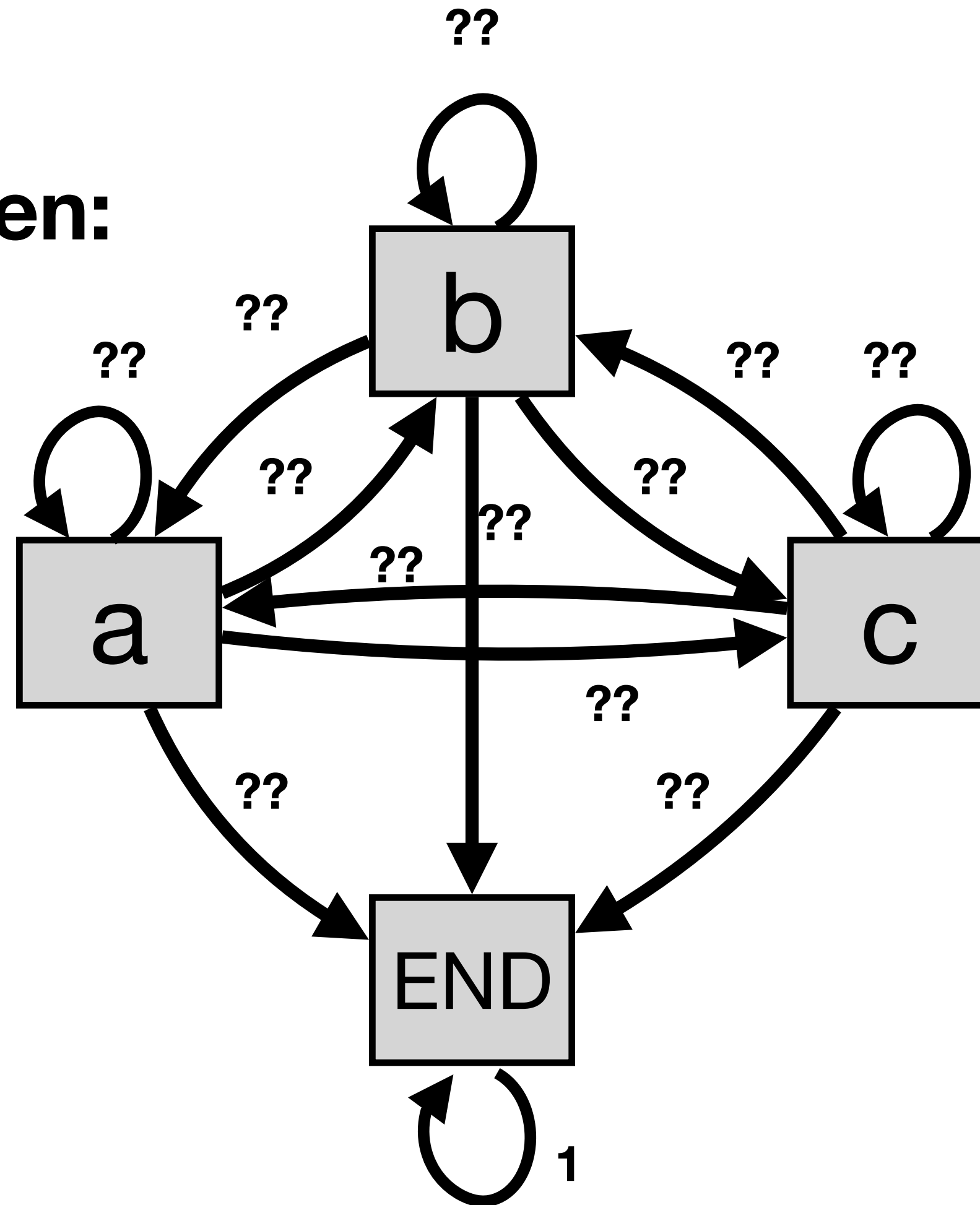
$$V^\pi(s) := G(s)/N(s) \text{ /* Update current estimate */}$$

Recall Our Example

Agent: 

Rewards??

States are given:



Actions are given:

$$A = \{L, R\}$$



Some policy π is given
(details not important now).

First-Visit MC Evaluation (Example)

Given: $S = \{a, b, c, \text{end}\}$, $A = \{L, R\}$, $\gamma = 1$

Sampled episodes (using given policy π):

$e_1 = a, R, 0, b, R, 10, c, L, 0, b, R, 0, c, R, 0, \text{end}$

$e_2 = a, R, 0, b, R, 10, c, L, 0, b, R, 10, a, L, 0, \text{end}$

After iteration 1:

After iteration 2:

Initialize: $G(s) = 0$, $N(s) = 0$ for all $s \in S$.

For $i = 1, \dots, N$:

Sample episode

$e_i := s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \dots, s_{i,T_i}$

For each time step $1 \leq t \leq T_i$:

If t is the first occurrence of state s in the episode e_i

$g_{i,t} := r_{i,t} + \gamma \cdot r_{i,t+1} + \gamma^2 \cdot r_{i,t+2} + \dots + \gamma^{T_i-t} \cdot r_{i,T_i}$

$N(s) := N(s) + 1$ /* Increment total visits counter */

$G(s) := G(s) + g_{i,1}$ /* Increment total return counter */

$V^\pi(s) := G(s)/N(s)$ /* Update current estimate */

First-Visit MC Evaluation (Example)

Given: $S = \{a, b, c, \text{end}\}$, $A = \{L, R\}$, $\gamma = 1$

Sampled episodes (using given policy π):

$e_1 = a, R, 0, b, R, 10, c, L, 0, b, R, 0, c, R, 0, \text{end}$

$e_2 = a, R, 0, b, R, 10, c, L, 0, b, R, 10, a, L, 0, \text{end}$

After iteration 1:

$G(a) = 10, G(b) = 10, G(c) = 0, G(\text{end}) = 0$

$N(a) = 1, N(b) = 1, N(c) = 1, N(\text{end}) = 1$

$V^\pi(a) = 10, V^\pi(b) = 10, V^\pi(c) = 0, V^\pi(\text{end}) = 0$

After iteration 2:

Initialize: $G(s) = 0, N(s) = 0$ for all $s \in S$.

For $i = 1, \dots, N$:

Sample episode

$e_i := s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \dots, s_{i,T_i}$

For each time step $1 \leq t \leq T_i$:

If t is the first occurrence of state s in the episode e_i

$g_{i,t} := r_{i,t} + \gamma \cdot r_{i,t+1} + \gamma^2 \cdot r_{i,t+2} + \dots + \gamma^{T_i-t} \cdot r_{i,T_i}$

$N(s) := N(s) + 1$ /* Increment total visits counter */

$G(s) := G(s) + g_{i,1}$ /* Increment total return counter */

$V^\pi(s) := G(s)/N(s)$ /* Update current estimate */

First-Visit MC Evaluation (Example)

Given: $S = \{a, b, c, \text{end}\}$, $A = \{L, R\}$, $\gamma = 1$

Sampled episodes (using given policy π):

$e_1 = a, R, 0, b, R, 10, c, L, 0, b, R, 0, c, R, 0, \text{end}$

$e_2 = a, R, 0, b, R, 10, c, L, 0, b, R, 10, a, L, 0, \text{end}$

After iteration 1:

$G(a) = 10, G(b) = 10, G(c) = 0, G(\text{end}) = 0$

$N(a) = 1, N(b) = 1, N(c) = 1, N(\text{end}) = 1$

$V^\pi(a) = 10, V^\pi(b) = 10, V^\pi(c) = 0, V^\pi(\text{end}) = 0$

After iteration 2:

$G(a) = 30, G(b) = 30, G(c) = 10, G(\text{end}) = 0$

$N(a) = 2, N(b) = 2, N(c) = 2, N(\text{end}) = 2$

$V^\pi(a) = 15, V^\pi(b) = 15, V^\pi(c) = 5, V^\pi(\text{end}) = 0$

Initialize: $G(s) = 0, N(s) = 0$ for all $s \in S$.

For $i = 1, \dots, N$:

Sample episode

$e_i := s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \dots, s_{i,T_i}$

For each time step $1 \leq t \leq T_i$:

If t is the first occurrence of state s in the episode e_i

$g_{i,t} := r_{i,t} + \gamma \cdot r_{i,t+1} + \gamma^2 \cdot r_{i,t+2} + \dots + \gamma^{T_i-t} \cdot r_{i,T_i}$

$N(s) := N(s) + 1$ /* Increment total visits counter */

$G(s) := G(s) + g_{i,1}$ /* Increment total return counter */

$V^\pi(s) := G(s)/N(s)$ /* Update current estimate */

Every-Visit Monte-Carlo Evaluation

Initialize: $G(s) = 0$, $N(s) = 0$ for all $s \in S$.

For $i = 1, \dots, N$:

Sample episode $e_i := s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \dots, s_{i,T_i}$.

For each time step $1 \leq t \leq T_i$:

~~If t is the first occurrence of state s in the episode e_i /* This was for first-visit MC */~~

s is the state visited at time t in the episode e_i

$$g_{i,t} := r_{i,t} + \gamma \cdot r_{i,t+1} + \gamma^2 \cdot r_{i,t+2} + \dots + \gamma^{T_i-t} \cdot r_{i,T_i}$$

$N(s) := N(s) + 1$ /* Increment total visits counter */

$G(s) := G(s) + g_{i,1}$ /* Increment total return counter */

$V^\pi(s) := G(s)/N(s)$ /* Update current estimate */

Every-Visit MC Evaluation (Example)

Given: $S = \{a, b, c, \text{end}\}$, $A = \{L, R\}$, $\gamma = 1$

Sampled episodes (using given policy π):

$e_1 = a, R, 0, b, R, 10, c, L, 0, b, R, 0, c, R, 0, \text{end}$

$e_2 = a, R, 0, b, R, 10, c, L, 0, b, R, 10, a, L, 0, \text{end}$

After iteration 1:

After iteration 2:

Initialize: $G(s) = 0$, $N(s) = 0$ for all $s \in S$.

For $i = 1, \dots, N$:

Sample episode

$e_i := s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \dots, s_{i,T_i}$

For each time step $1 \leq t \leq T_i$:

s is the state visited at time t in the episode e_i

$g_{i,t} := r_{i,t} + \gamma \cdot r_{i,t+1} + \gamma^2 \cdot r_{i,t+2} + \dots + \gamma^{T_i-t} \cdot r_{i,T_i}$

$N(s) := N(s) + 1$ /* Increment total visits counter */

$G(s) := G(s) + g_{i,1}$ /* Increment total return counter */

$V^\pi(s) := G(s)/N(s)$ /* Update current estimate */

Every-Visit MC Evaluation (Example)

Given: $S = \{a, b, c, \text{end}\}$, $A = \{L, R\}$, $\gamma = 1$

Sampled episodes (using given policy π):

$e_1 = a, R, 0, b, R, 10, c, L, 0, b, R, 0, c, R, 0, \text{end}$

$e_2 = a, R, 0, b, R, 10, c, L, 0, b, R, 10, a, L, 0, \text{end}$

After iteration 1:

$G(a) = 10, G(b) = 10, G(c) = 0, G(\text{end}) = 0$

$N(a) = 1, N(b) = 2, N(c) = 2, N(\text{end}) = 1$

$V^\pi(a) = 10, V^\pi(b) = 5, V^\pi(c) = 0, V^\pi(\text{end}) = 0$

After iteration 2:

Initialize: $G(s) = 0, N(s) = 0$ for all $s \in S$.

For $i = 1, \dots, N$:

Sample episode

$e_i := s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \dots, s_{i,T_i}$

For each time step $1 \leq t \leq T_i$:

s is the state visited at time t in the episode e_i

$g_{i,t} := r_{i,t} + \gamma \cdot r_{i,t+1} + \gamma^2 \cdot r_{i,t+2} + \dots + \gamma^{T_i-t} \cdot r_{i,T_i}$

$N(s) := N(s) + 1$ /* Increment total visits counter */

$G(s) := G(s) + g_{i,t}$ /* Increment total return counter */

$V^\pi(s) := G(s)/N(s)$ /* Update current estimate */

Every-Visit MC Evaluation (Example)

Given: $S = \{a, b, c, \text{end}\}$, $A = \{L, R\}$, $\gamma = 1$

Sampled episodes (using given policy π):

$e_1 = a, R, 0, b, R, 10, c, L, 0, b, R, 0, c, R, 0, \text{end}$

$e_2 = a, R, 0, b, R, 10, c, L, 0, b, R, 10, a, L, 0, \text{end}$

After iteration 1:

$G(a) = 10, G(b) = 10, G(c) = 0, G(\text{end}) = 0$

$N(a) = 1, N(b) = 2, N(c) = 2, N(\text{end}) = 1$

$V^\pi(a) = 10, V^\pi(b) = 5, V^\pi(c) = 0, V^\pi(\text{end}) = 0$

After iteration 2:

$G(a) = 30, G(b) = 40, G(c) = 10, G(\text{end}) = 0$

$N(a) = 3, N(b) = 4, N(c) = 3, N(\text{end}) = 2$

$V^\pi(a) = 10, V^\pi(b) = 10, V^\pi(c) = \frac{10}{3}, V^\pi(\text{end}) = 0$

Initialize: $G(s) = 0, N(s) = 0$ for all $s \in S$.

For $i = 1, \dots, N$:

Sample episode

$e_i := s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \dots, s_{i,T_i}$

For each time step $1 \leq t \leq T_i$:

s is the state visited at time t in the episode e_i

$g_{i,t} := r_{i,t} + \gamma \cdot r_{i,t+1} + \gamma^2 \cdot r_{i,t+2} + \dots + \gamma^{T_i-t} \cdot r_{i,T_i}$

$N(s) := N(s) + 1$ /* Increment total visits counter */

$G(s) := G(s) + g_{i,t}$ /* Increment total return counter */

$V^\pi(s) := G(s)/N(s)$ /* Update current estimate */

Statistical Properties (1/7)

- First-visit MC Policy Evaluation is **unbiased** (and hence also consistent) estimator.
- Every-visit MC Policy Evaluation is a **biased** but **consistent** estimator, which often has better MSE.

Statistical Properties (2/7)

First-visit MC Policy Evaluation is **unbiased** (and hence also consistent) estimator.

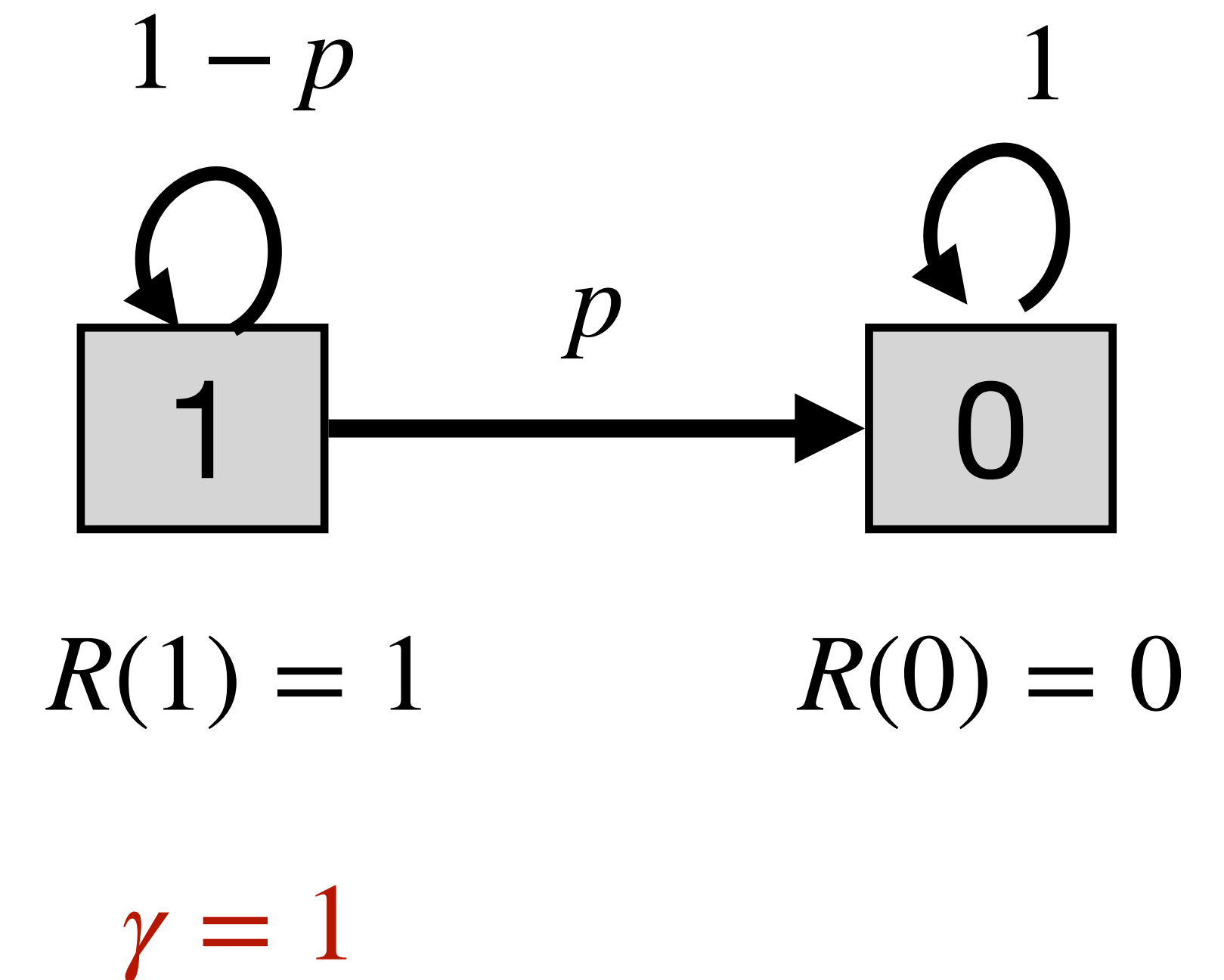
Proof Sketch:

Assuming Markov property, the first occurrence* of the state s at time t together with the subsequence starting at t gives us an unbiased estimate of the return starting from s (this is practically from definition), i.e., $\mathbb{E}[G_t^\pi | X_t = s]$, which is by definition equal to $V^\pi(s)$. First-visit MC averages such independent samples from different episodes (different episodes \Rightarrow independence).

**Do you see why we cannot take, e.g., the last occurrence? Hint: Are subsequences starting with the last occurrence of s special in some way?*

Statistical Properties (3/7)

- Every-visit MC Policy Evaluation is a **biased** but **consistent** estimator, which often has **better MSE**.
- **Example (Showing that it is biased):**



Statistical Properties (4/7)

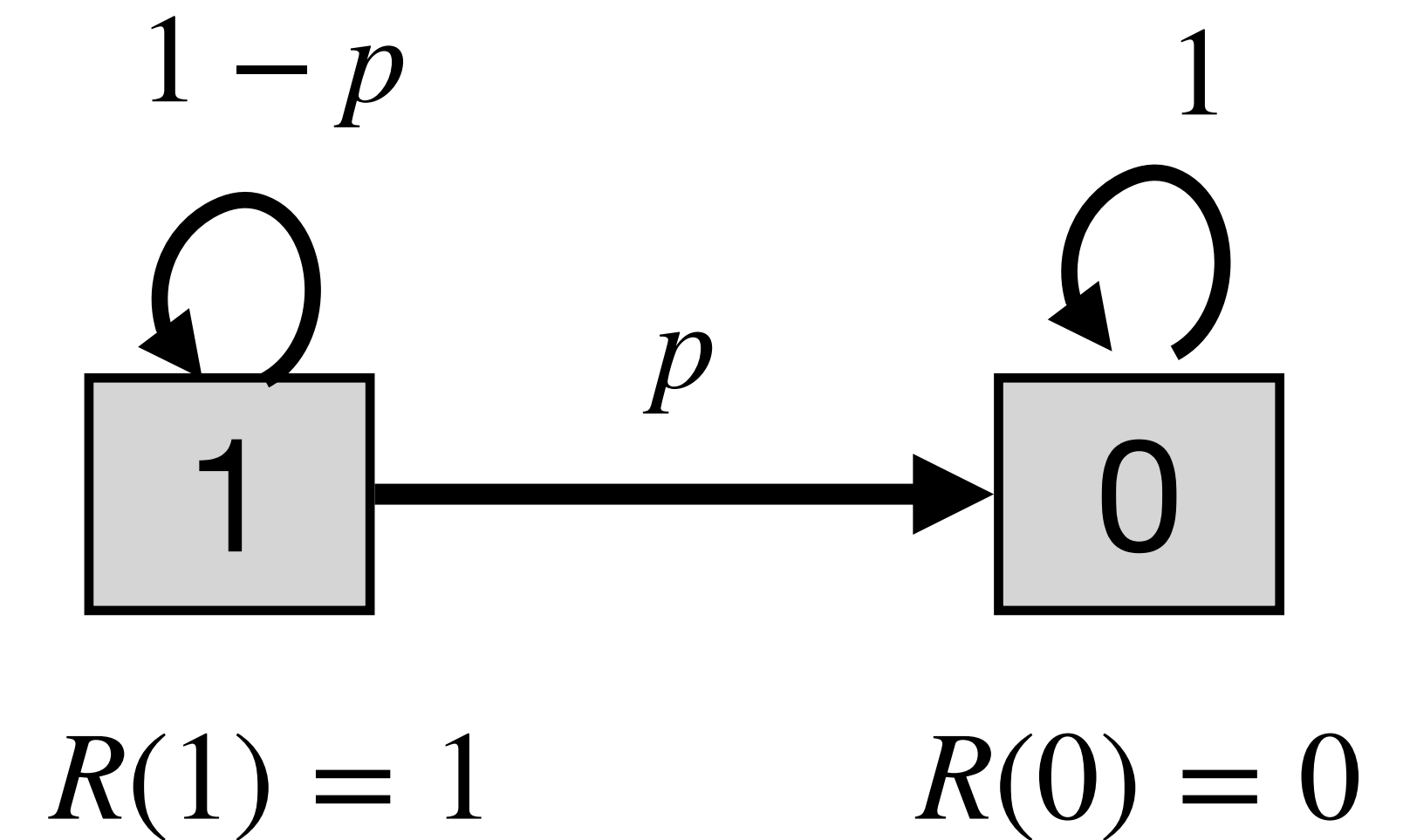
- Every-visit MC Policy Evaluation is a **biased** but **consistent** estimator, which often has **better MSE**.

- **Example (Showing that it is biased):**

- Computing V explicitly using Bellman equation:

$$V(1) = 1 + (1 - p) \cdot V(1) + p \cdot 0$$

$$\text{Hence, } V(1) = \frac{1}{p}.$$



$$\gamma = 1$$

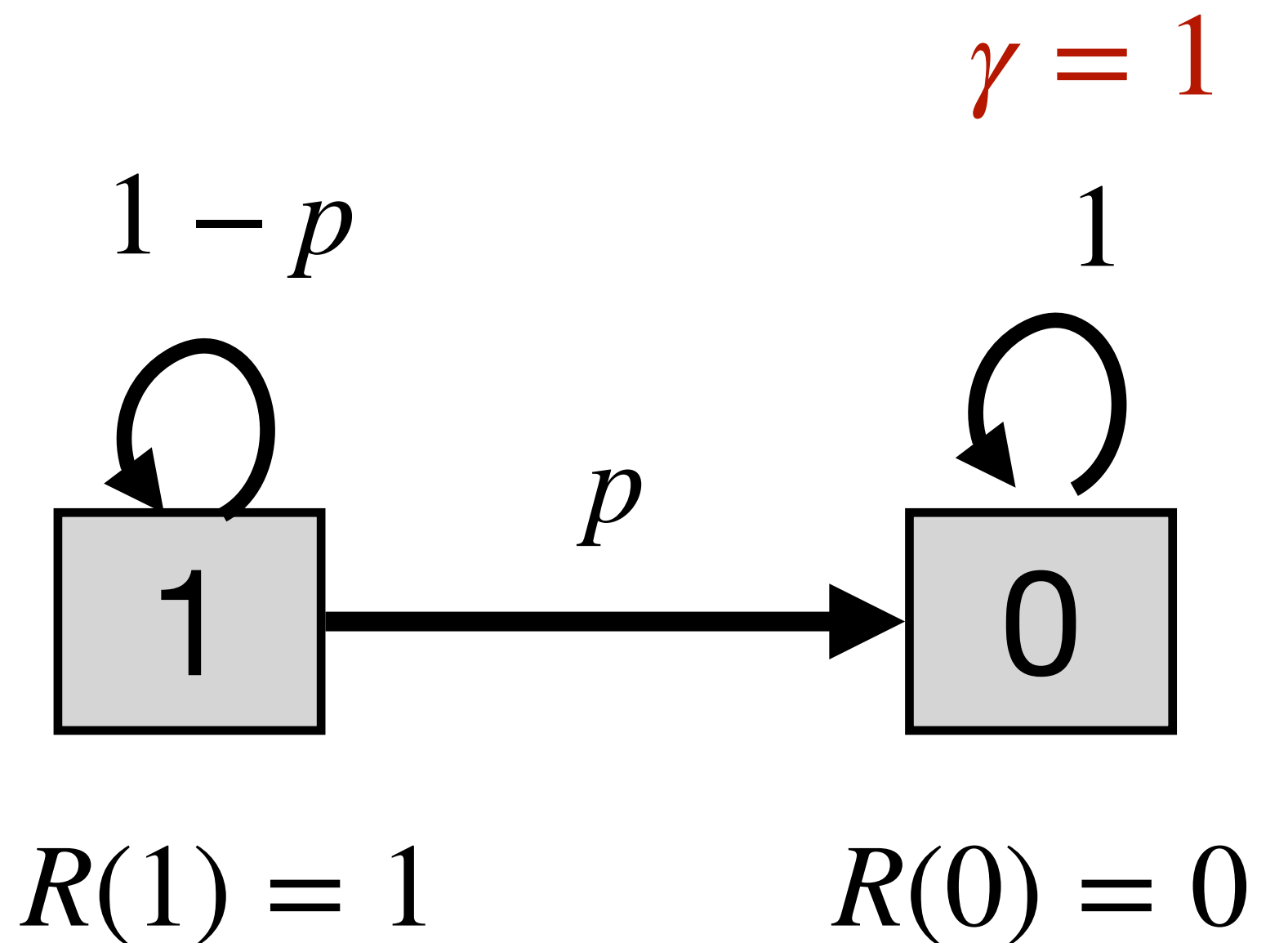
Statistical Properties (5/7)

- Every-visit MC Policy Evaluation is a **biased** but **consistent** estimator, which often has **better MSE**.

- **Example (Showing that it is biased):**

- Exact answer: $V(1) = \frac{1}{p}$.

- First-Visit MC:



$$\mathbb{E}[\hat{V}_{FV}] = p + 2(p-1)p + 3(p-1)^2p + \dots = p \sum_{n=0}^{\infty} (n+1) \cdot (1-p)^n = p \cdot \frac{1}{p^2} = \frac{1}{p}$$

UNBIASED

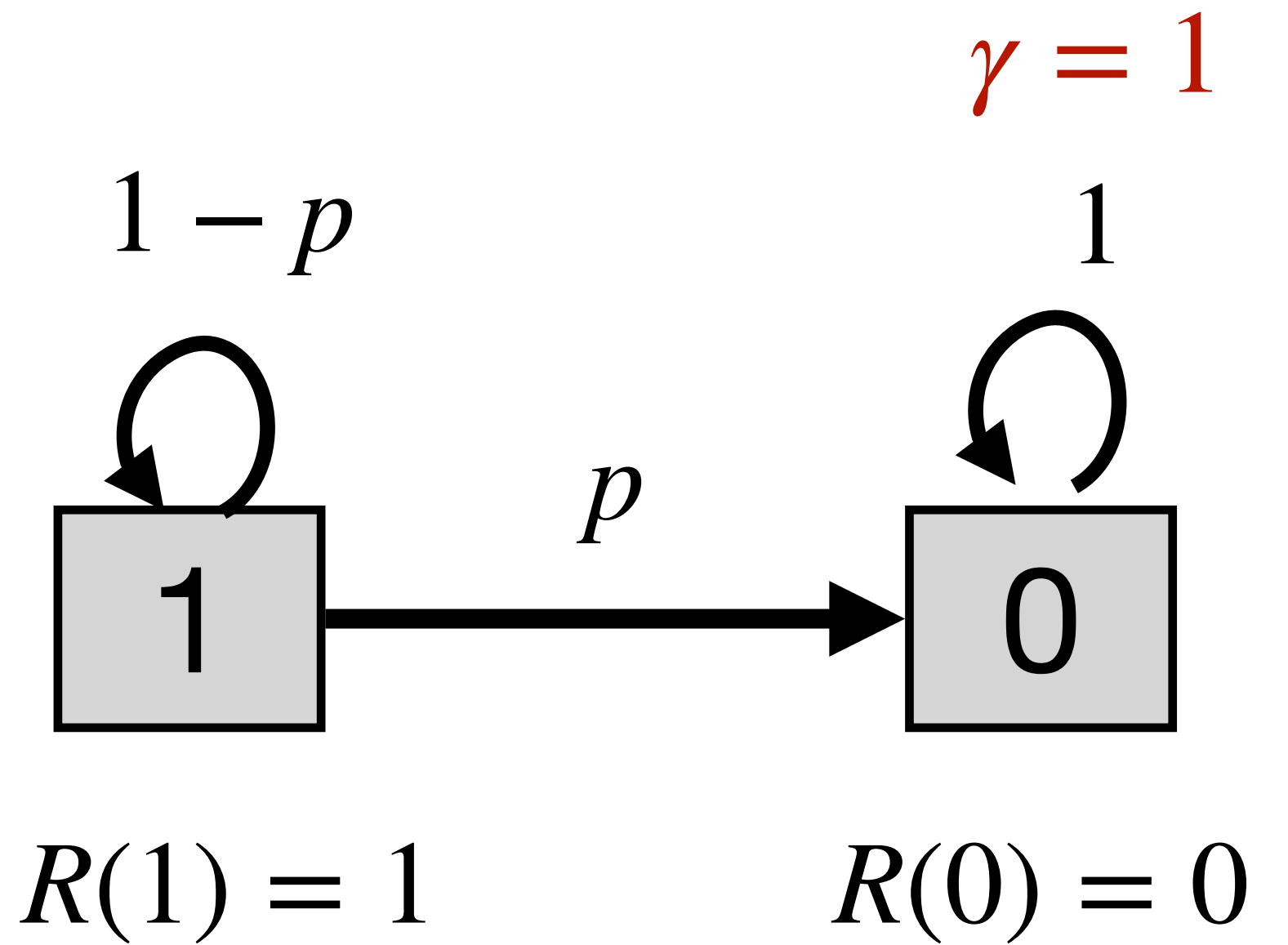
Statistical Properties (6/7)

- Every-visit MC Policy Evaluation is a **biased** but **consistent** estimator, which often has **better MSE**.

- **Example (Showing that it is biased):**

- Exact answer: $V(1) = \frac{1}{p}$.

- Every-Visit MC (Bias):



$$\mathbb{E}[\hat{V}_{EV}] = p + \frac{3}{2}(p-1)p + 2(p-1)^2p + \dots = p \sum_{n=0}^{\infty} \frac{n+2}{2} \cdot (1-p)^n = p \cdot \frac{p+1}{p^2} = \frac{p+1}{2p} \neq \frac{1}{p}$$

BIASED

Statistical Properties (7/7)

- Every-visit MC Policy Evaluation is a **biased** but **consistent** estimator, which often has **better MSE**.

- **Example (Showing that it is biased):**

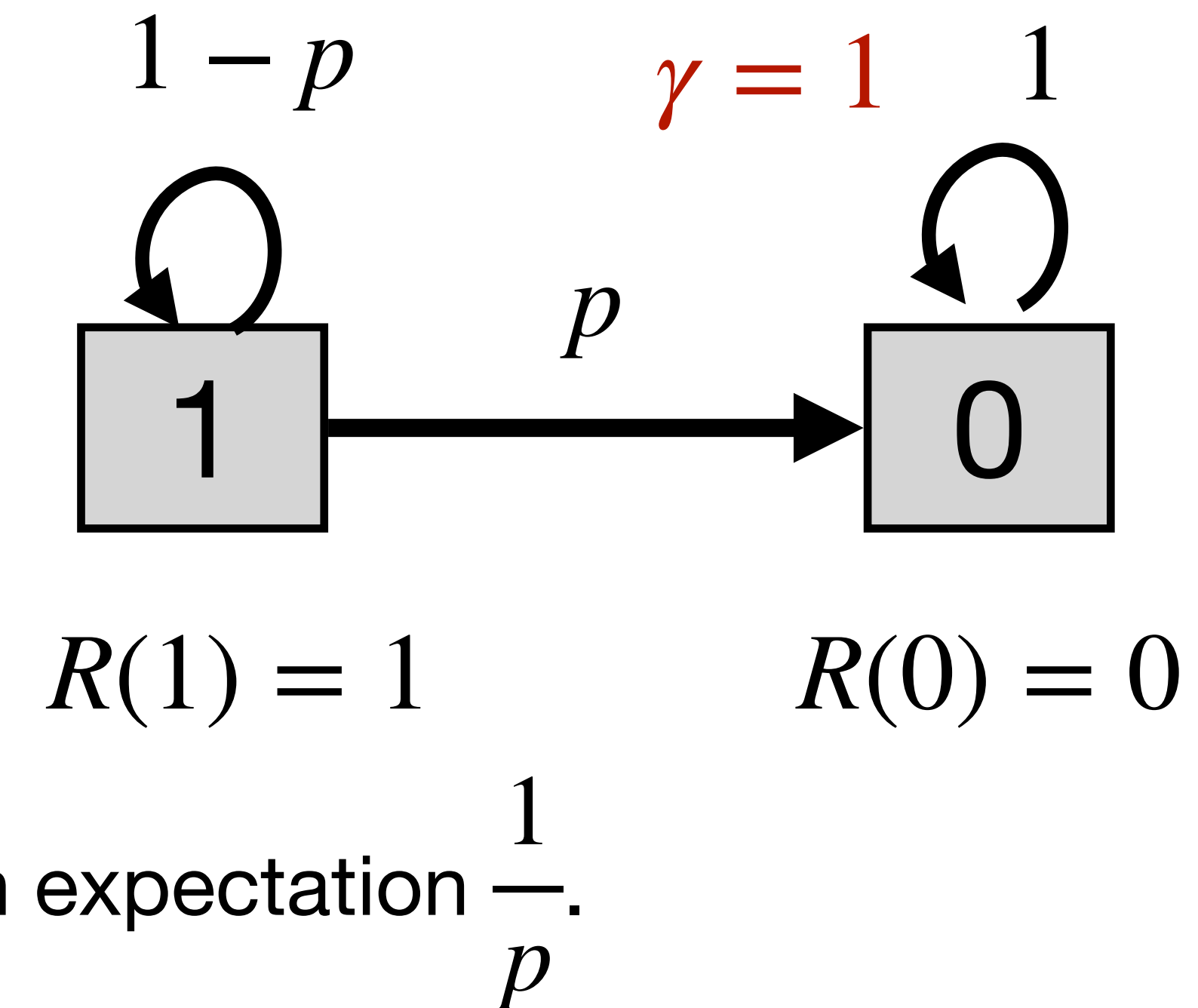
- Exact answer: $V(1) = \frac{1}{p}$.

- Every-Visit MC (**Consistency**):

$$\hat{V}_{EV} = \frac{T + 1}{2} \text{ where } T \text{ is a geometrically distributed r.v. with expectation } \frac{1}{p}.$$

Averaging estimators over n independent episodes, one can show with a bit of algebraic

manipulations that $P \left[\left| \hat{V}_n - \frac{1}{p} \right| < \varepsilon \right] = 1$ for all $0 < \varepsilon$.



Consistent

Incremental Monte-Carlo Evaluation

Initialize: $G(s) = 0$, $N(s) = 0$ for all $s \in S$.

For $i = 1, \dots, N$:

Sample episode $e_i := s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \dots, s_{i,T_i}$.

For each time step $1 \leq t \leq T_i$:

s is the state visited at time t in the episode e_i

$$g_{i,t} := r_{i,t} + \gamma \cdot r_{i,t+1} + \gamma^2 \cdot r_{i,t+2} + \dots + \gamma^{T_i-t} \cdot r_{i,T_i}$$

$$N(s) := N(s) + 1 \text{ /* Increment total visits counter */}$$

$$V^\pi(s) := V^\pi(s) + \alpha \cdot (g_{i,t} - V^\pi(s)) \text{ /* Update value function */}$$

Special case: When we use $\alpha = \frac{1}{N(s)}$ then the resulting incremental MC becomes equivalent to every-visit MC.

Summary (So Far)

- **MC Methods:**
 - Try to estimate $V^\pi(s) = \mathbb{E}[G_t^\pi | X_t = s]$ directly as an average over sampled episodes (which is also why they need the episodic settings).
 - They do not use the Markov assumption!
 - Converge to the true mean.
 - Can have high variance and some of them are also biased (first-visit MC is onw which is not biased).

Part 4: Temporal Difference Learning

(We are still dealing with policy evaluation)

Temporal Difference Learning: A Teaser

- **TD learning** combines Monte-Carlo estimation and dynamic programming ideas.
- **TD learning** can be used both in episodic and infinite-horizon non-episodic settings,
- **TD learning** updates estimates of V^π continually, after every consecutive tuple *state-action-reward-state* (therefore we do not need to wait till the end of an episode).

....

TD-Learning: Basic Idea

Recall: $g_{i,t} := r_{i,t} + \gamma \cdot r_{i,t+1} + \gamma^2 \cdot r_{i,t+2} + \dots + \gamma^{T_i-t} \cdot r_{i,T_i}$

Incremental MC:

$$V^\pi(s) := V^\pi(s) + \alpha \cdot (g_{i,t} - V^\pi(s)).$$

Temporal Difference Learning:

$$V^\pi(s_t) := V^\pi(s_t) + \alpha (r_{i,t} + \gamma \cdot V^\pi(s_{t+1}) - V^\pi(s_t))$$

\approx

TD-Learning: Relationship to Bellman Backup

Recall: $g_{i,t} := r_{i,t} + \gamma \cdot r_{i,t+1} + \gamma^2 \cdot r_{i,t+2} + \dots + \gamma^{T_i-t} \cdot r_{i,T_i}$

Bellman equation:

$$V_{k+1}^\pi(s) := R(s, \pi(s)) + \gamma \cdot \sum_{s' \in \mathcal{S}} P(s' | s, \pi(s)) \cdot V_k(s')$$

Expectation

Temporal Difference Learning:

$$\begin{aligned} V^\pi(s_t) &:= V^\pi(s_t) + \alpha(r_{i,t} + \gamma \cdot V^\pi(s_{t+1}) - V^\pi(s_t)) \\ &= (1 - \alpha) \cdot V^\pi(s_t) + \alpha \cdot (r_{i,t} + \gamma \cdot V^\pi(s_{t+1})) \end{aligned}$$

Sample

TD-Learning: Pseudocode

Initialize: $V^\pi(s) = 0$ for all $s \in \mathcal{S}$

Loop:

Sample tuple (s_t, a_t, r_t, s_{t+1}) .

Update $V^\pi(s_t) := V^\pi(s_t) + \alpha \cdot \underbrace{(r_{i,t} + \gamma \cdot V^\pi(s_{t+1}) - V^\pi(s_t))}_{\text{TD target}}$

TD-Learning: Example

Initialize: $V^\pi(s) = 0$ for all $s \in S$

Loop:

Sample tuple (s_t, a_t, r_t, s_{t+1}) .

Update $V^\pi(s_t) := V^\pi(s_t) + \alpha \cdot \underbrace{(r_{i,t} + \gamma \cdot V^\pi(s_{t+1}) - V^\pi(s_t))}_{\text{TD target}}$

$$\alpha = 0.5, \gamma = 1$$

$e_1 = a, R, 0, b, R, 10, c, L, 0, b, R, 0, c, R, 0, \text{end}$

Iteration 1: $V^\pi(a) := 0,$

Iteration 2: $V^\pi(b) := 5,$

Iteration 3: $V^\pi(c) := 0.5(0 + 5) = 2.5,$

Iteration 4: $V^\pi(b) := 5 + 0.5 \cdot (0 + 2.5 - 5) = 3.75,$

Iteration 5: $V^\pi(c) := 2.5 + 0.5 \cdot (0 + 0 - 2.5) = 1.25.$

What About the α 's?

- One thing we can do is to have α depend on the number of iterations so far, i.e., we can have α_k instead of just α .
- Convergence is guaranteed when α_k 's satisfy the following conditions (follows from Robbins-Munro algorithm):

$$\sum_{k=1}^{\infty} \alpha_k = \infty, \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

- A sequence which satisfies the above conditions is, e.g., $\alpha_k = \frac{1}{k}$. However, in practice, similar sequences do not have to converge very fast...
- *Note: It was also proved by Sutton (1988) that, for tabular MDPs, there always exists some small enough learning rate α such that TD converges but this result is not very practical.*

Policy Evaluation: Summary

	DPCE	MC	TD
Can use w/out access to true MDP models	X	X	X
Usable in continuing (non-episodic) setting	X		X
Assumes Markov process	X		X
Converges to true value in limit ³	X	X	X
Unbiased estimate of value		X	

of certainty

- DPCE = Dynamic Programming w/certainty equivalence estimates, MC = Monte Carlo, TD = Temporal Difference

Table from slides by Prof. Emma Brunskill